# Where does all the data come from?

Geoffrey Fox (Indiana University, USA), Tony Hey (Microsoft Research, USA) and Anne Trefethen (Oxford University, UK)

### Abstract

….

## Introduction

Data without context is of little or no value. It matters where data has come from and how it has been processed up to that point. This is the ‘data provenance’ that must be associated with the data for it to be usable by other researchers. Digital information is being generated in large quantities each day and, depending on the source, the data comes with a variety of issues - ranging from questions about the appropriateness of the metadata and semantics to describe it, to the integrity and completeness of the data. Finally, all data comes with a cost for keeping it and pragmatic choices must be made about what data to keep and for how long - since it is impractical to retain all the data we are now generating. In some fields, such as particle physics and astronomy, the raw data rates are now so large that only a selection of the data can be retained for future analysis.

The preparation of data for possible re-use and preservation is the process of ‘data curation’. This includes such things as data-cleansing, to check the integrity of the data, and the adding of metadata – data about the data – to document what the data is, how the data has been collected and what format has been used and so on, and perhaps adding higher-level annotations or semantic information by using ontologies or community-agreed vocabularies. All of these things are necessary to facilitate not only the ‘findability’ of data by search engines but also to allow the possibility of re-use and the creation of meaningful ‘scientific mashups’ of different data sets. Unfortunately, scientists are only human and the process of adding value to their data sets for others to use often receives scant reward. It is therefore not surprising that, all too often, the data curation and preservation procedures used will be imperfect or inadequate and that some important data will end up effectively being lost to future researchers. At least a part of our future occupation will likely to be something like a ‘digital archaeologist’ - trying to make sense of old data by piecing together the fragments of an inadequately documented historic past.

An interesting cautionary tale is that of the digital Domesday Book project in the UK [1]. After William the Conqueror won the Battle of Hastings and took control of what later became the kingdom of England, he decided to take a census of his new realm. The result was the Domesday book completed in 1086 which can still be seen in the National Archives in London. In 1986, 900 years after the original census, the BBC produced a television program to celebrate this anniversary. One result of this project was an interactive video documentary implemented for the BBC microcomputer. At the launch of the UK Digital Preservation Coalition in London in 2002, broadcaster Lloyd Grossman called attention to the danger that the rapid evolution of computer media and recording formats would lead to the irretrievable loss of valuable historic records, and he gave the Domesday Book video-disks as an example. His remark sparked a heroic effort by many people to rescue the digital images, text and video of the BBC Domesday interactive-video project. Eventually,after a lot of work by some exceptionally dedicated individuals, a version was produced in modern formats and that works on a modern PC.

Data born digital comes in many sizes and shapes and from a vast range of sources – even in our daily lives we create a massive amount of digital information about ourselves and our lives. We do this by shopping with credit cards, using on-line systems, social networking, sharing videos, capturing traffic flow, measuring pollution, using security cameras, and so on. In addition, in healthcare we now routinely create digital medical images and some sort of ‘electronic health record’ or HER. In this chapter, after looking briefly at the explosion of digital data and devices that are impacting our daily lives, we shall focus on some examples of the scientific data deluge. We anticipate that some of the tools, techniques and visualizations that are now being used by scientists to explore and manage their challenging data sets will find their way into business and everyday life – just as was done with creation of the Web by Tim Berners-Lee and the particle physics community. In our focus on scientific data, it is important not to ignore the digital aspects of the social sciences and the humanities which, although physical artifacts still play an important role in research, are becoming increasingly important.

## An Explosion of Data in our lives

A recent report by McKinsey Global Institute [2] reports that there are 30 billion pieces of content shared on Facebook every month, that the US Library of Congress had collected 235 terabytes of data by April 2011 and that 15 out of 17 sectors in the US have more data stored per company than stored by the US Library of Congress. The future only sees this increasing, with a projected growth of 40% in global data generated by commerce and individuals per year. The McKinsey report estimates that globally, enterprises stored more than 7 exabytes of new data on disk drives in 2010, comprising data generated through interactions with a customer base and data supporting the provision of services through the Internet. Individuals stored a similarly impressive amount - more than 6 exabytes of new data on home and hand-held devices.

The report also notes that there are over 30 million networked sensors deployed in the transportation, industrial, retail, and utilities sectors and that this number is increasing by more than 30% per year. We will return to the issue of sensor data below as we look in more detail at the Smart Grid and oceanographic examples of data collection.

It is a truism to say that the Internet has changed everything - today our lives are often as much digital as physical. Our collaborations and friendships are as likely to be virtual as being based on real, physical, face-to-face meetings and the management of our lives – be it banking, house utilities, health, or car insurance – is increasingly dominated by networked systems and on-line commerce.

The 2010 Verisign report [3] indicates that by the end of 2010 there will be over 202 million domain name registrations - an increase of 7% over the previous year. It notes that the largest Top Level Domains or TLDs in terms of base size were, in order, .com, .de (Germany), .net, .uk (United Kingdom), .org, .info, .cn, .nl (Netherlands), .eu (European Union) and .ru (Russian Federation). Figure (1.1) below shows the breakdown.

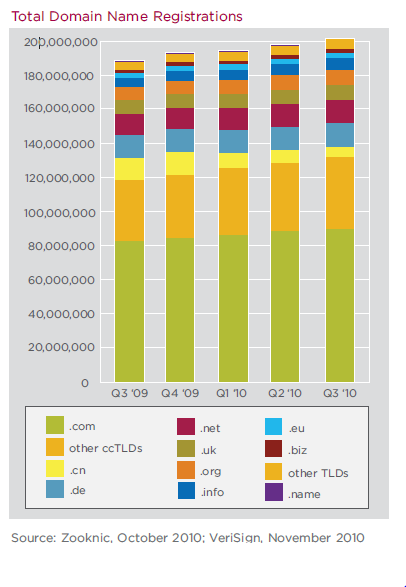


Figure 1.1: Breakdown of Top Level Domains

Using this information, TNW (The Next Web) [4] estimate the number of pages on the web to be between 42 billion and 121 billion, a 21 % increase from 2008. Verisign also report that in the last decade, the number of Internet users has increased by 500%. It is notable, however, that this growth is not homogenous globally. Certain international regions are exploding in their usage of the Internet. For example, a decade ago, Africa had less than 5 million Internet users: it now has more than 100 million. The report also estimates that in 2010 less than 40% of Internet users were English speaking.

Social use of the Internet is generating content constantly and in ever increasing amounts. Search Engine Watch [5] includes some interesting statistics on YouTube, the video sharing community site.In 2010, more than 13 million hours of video were uploaded to the site; in two months (60 days) more video was uploaded than had been created in six decades by the three major networks (ABC, CBS, NBC).By May 2011, more than 48 hours of video were being uploaded per minute, and YouTube had surpassed 3 billion views per day.

The GreenPeace report “How clean is your data?” [6] reports that 1.2 Zettabytes of digital information has been generated by tweets, by Facebook where over 30 billion pieces of content are shared each month, emails, YouTube and other social data transfers. The use of these social network and related tools is beautifully illustrated below by JESS3[[1]](#footnote-1). [Fig. 1.2 ]

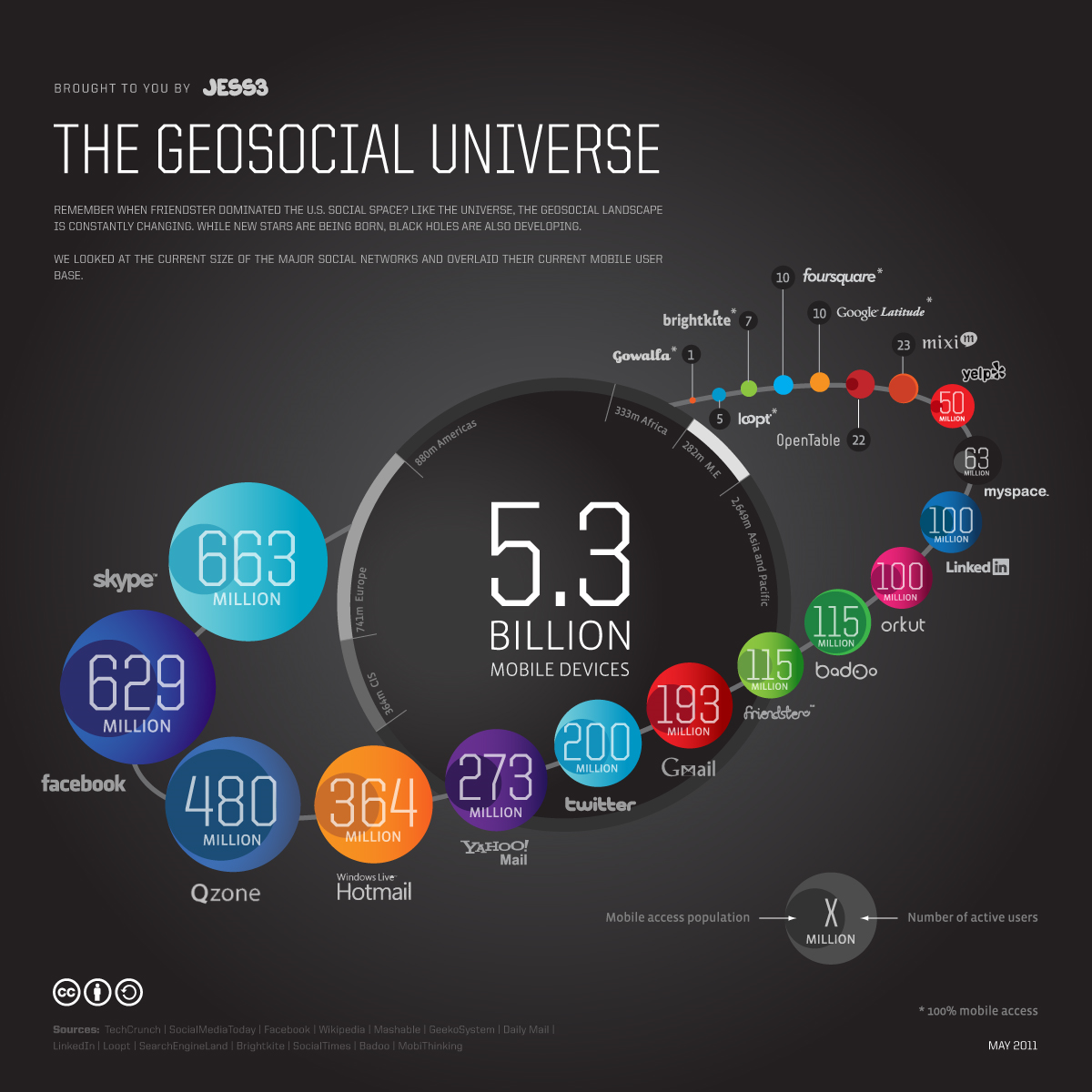


Figure 1.2: The Geosocial Universe

Flickr, the photo sharing site now hosts over 4 billion images. These images are generally family photographs, holiday snaps and the like. However, the increasing demand for digital image storage is a growing concern as we turn to healthcare and the medical images that are now a standard part of our healthcare systems. Medical images are created in many forms and for a broad range of diagnostic reasons. They include magnetic resonance imaging (MRI), digital mammography, positron emission tomography (PET), and X-ray computed tomography (CT). Together, these images amount to over 20 thousand terabytes of data: a single CT study consists of sixty-four 512 x 512 x 16 bit images and can correspond to as much as tens or even hundreds of megabytes. In the case of mammography, the size of digital data collected is approximately 200 megabytes per examination.

Ninety-six percent of radiology practices in the USA are filmless and Table 1.1 below illustrates the annual volume of data across the types of diagnostic imaging [7]; this does not include cardiology which would take the total to over 1,000,000,000GB.

Table 1.1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Modality** | **Part B non HMO** | **All Medicare** | **All Population** | **Per 1000 persons** | **Ave study size (GB)** | **Total annual data generated in GB** |
| CT | 22 million | 29 million | 87 million | 287 | 0.25 | **21,750,000** |
| MR | 7 million | 9 million | 26 million | 86 | 0.2 | **5,200,000** |
| Ultrasound | 40 million | 53 million | 159 million | 522 | 0.1 | **15,900,000** |
| Interventional | 10 million | 13 million | 40 million | 131 | 0.2 | **8,000,000** |
| Nuclear Medicine | 10 million | 14 million | 41 million | 135 | 0.1 | **4,100,000** |
| PET | 1 million | 1 million | 2 million | 8 | 0.1 | **200,000** |
| Xray, total incl. mammography | 84 million | 111 million | 332 million | 1,091 | 0.04 | **13,280,000** |
|  |  |  |  |  |  |  |
| All Diagnostic Radiology | 174 million | 229 million | 687 million | 2,259 | 0.1 | **68,700,000** |

The image repositories required to store medical images are more complex than those required to store photos on Flickr – the data are more complex, there are often diverse user access requirements, and often the need to search and analyze the data in collections, according to particulars of an individual or to a specific disease type. The National Cancer Institute maintains a survey of biomedical imaging archives [7] but the terabytes listed there are only the tip of the iceberg.

There are also other considerations addressing the management of medical images. In the US, providers must meet HIPAA (Health Insurance Portability and Accountability Act) regulations that require a data back-up plan and disaster recovery plans. BridgeHead Software, who describe themselves as a Healthcare Storage Virtualization company, found the top IT spending priorities of hospitals for 2010 were disaster recovery (44%), Picture Archiving and Communication Systems or PACS (38%) and digitizing paper records (35%) [8]. In addition, each state in the US has its own medical record retention rules, typically for a minimum of seven years.

Our actions in everyday life are captured by sensors of many kinds – pollution sensors in cities, ctv throughout many public places, and increasingly, by new technologies such as Smart Grids for electricity [9]. Smart Grids are a relatively new approach to energy management. A Smart Grid is generally an energy network that incorporates information technology to allow real-time management of the energy generation and distribution using two-way communication between generators and end-users. In the UK, the Department of Energy and Climate Change have set about rolling out smart meters across the UK [10] and its anticipated that by 2019, 50 million smart meters will have been installed. The likely data generated and collected from these meters includes 12 readings per hour at 50 bytes per reading which, together with the system monitoring tools, will create Gbytes of data each. The challenges for the Smart Grid are not only technical – how to store and analyse such data – but also raise issues of security and policy. How will the data be kept private? How long will it be kept for? How much data should be kept locally versus centrally. And of course the telecommunications networks required to implement such capabilities are not generally readily available. However, the move to Smart Grids is global and as micro-generation of electricity increases through local renewable energy schemes, the requirements for Smart Grid technologies will only increase [11].Finally, an IDC report tiltled ‘The 2011 Digital Universe Study: Extracting Value from Chaos’ estimates that in 2011 the total volume of information created and replicated ‘will surpass 1.8 Zettabytes (1.8 trillion gigabytes) - growing by a factor of 9 in just five years’ [12]. The report estimates that 75% of this digital content will be created by individuals. The report also estimates that the number of files is growing even faster than the information itself and attributes this rapid growth to the increasing number of embedded systems now generating data.

## Research Data

Scientific research creates its data from observations, experimentation and simulation, and of course the outputs of such research include files and databases as well as publications in digital form. These textual publications increasingly need to be linked to the underlying primary data. In this chapter we will consider some examples and look at some of the issues that we must face in dealing with the data that is being generated.

## Observations

### Astronomy

#### The Square Kilometre Array (SKA)

At the present time hundreds of astronomers, computer scientists and technology engineers are designing the next generation radio telescope - the Square Kilometre Array or SKA [13]. It is anticipated that construction of the first phase of the telescope will begin in 2016 with the full telescope completed and in operation by 2022. The decision where to locate the SKA will be made in 2012 and it is likely be located in either Australia or South Africa, in a desert so as to have little or no interference, but will be a collaborative effort involving over 50 groups in 19 countries.

The present design [14,15,16] has a combination of aperture arrays in the core and up to 3000 phased array feeds on dishes giving a collecting area of approximately one square kilometer, with receptors extending out to a distance of 3000km from the centre of the telescope (Figure 1.3). The SKA will have a sensitivity of more than 50 times that of existing telescopes, and 10,000 times the survey speed. It is intended to provide data to answer fundamental questions about gravitation and magnetism, galaxy formation and even the question of life on other planets. The design of the SKA is developing through studies based on the science requirements, and on a number of SKA ‘Pathfinder’ projects that provide experience of design options, and technology capability considerations.

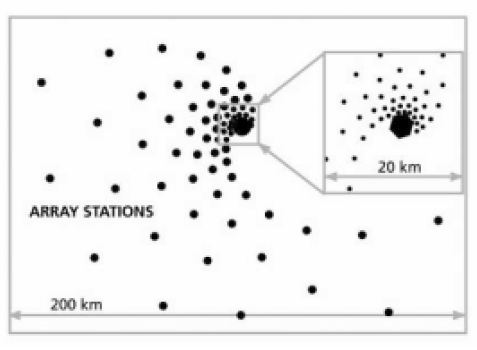


Figure 1.: Possible Configuration of SKA and artist's impression of the SKA core

The SKA provides an enormous information technology challenge, with a typical data rate from each dish antenna on the order of 100Gbs-1 aggregating to over 100Tbs-1 [17] and need for exaflop-scale computation [18] for post-processing. The IT infrastructure required to support the science at the SKA will range from real-time capability to transport and analyse the data at these high-data rates together with the capacity to store and “publish” the data for later analysis and interpretation by the global astrophysics community. The computational systems will likely range from specifically designed FPGA-like units to exascale computing systems and Cloud data centres. The communications infrastructure will range from high bandwidth intra-chip and inter-chip connectionson a high-performance computing system, to trans-oceanographic data transmission optical fibers supporting data rates of at least 100Gbs-1. The SKA will succeed or not depending on both the physical implementation of the telescope design and the software infrastructure that will enable it. The software infrastructure required to realise this information technology challenge is itself has been identified as > 2000 person year task [19] but even this may not take full account of the complexity of the task.

Figure : Computational requirements for data analysis from SKA

#### The Large Synoptic Survey Telescope (LSST)

The Large Synoptic Survey Telescope (LSST) [20] is the most ambitious survey currently planned in the optical part of the electromagnetic spectrum. The LSST is driven by four main science themes: probing dark energy and dark matter, taking an inventory of the Solar System, exploring the transient optical sky, and mapping the Milky Way. It will be a large, wide-field ground-based telescope designed to obtain multiple images covering the sky that is visible from Cerro Pachón in Northern Chile. The current LSST design has an 8.4m (6.7m effective) primary mirror, a 9.6 deg2 field of view, and a 3.2 Gigapixel camera. This will allow about 10,000 square degrees of sky to be covered using pairs of 15-second exposures twice per night, every three nights on average. The system is designed to yield high image quality as well as high astrometric and photometric accuracy. The total survey area will include 30,000 deg2 and will be imaged multiple times in six bands covering the wavelength range 320—1050 nm. The project is scheduled to begin the regular survey operations before the end of this decade. About 90% of the observing time will be devoted to a deep-wide-fast survey mode which will uniformly observe a 18,000 deg2 region about 1000 times (summed over all six bands) during the anticipated 10 years of operations. These data will result in databases including 10 billion galaxies and a similar number of stars, and will serve the majority of the primary science programs [21].

In terms of numbers, LSST will handle 15 terabytes of raw scientific image data each night. The final image data archive is estimated to have around 200 petabytes of data. It is estimated that the project will require a sustained petaflop/s computing capability and, of course, significant local processing power in Chile and very high bandwidth connection to the US archive site. The project plans to use SciDB, a novel open source database system that is optimized for scientific data management of ‘big data’ and for ‘big analytics’ [22].

## Earth and Ocean Observations

#### Earth Observation Data

NASA’s Earth Observing System Data and Information System [23] manages and distributes data products through its Distributed Active Archive Centers or DAACs. Each DAAC processes, archives, documents, and distributes data from NASA’s past and current research satellites and field programs. Each center serves one or more specific Earth science sub-discipline and provides appropriate data products, data information and services for its community. In Europe, the European Space Agency ESA [24] plays a similar role to EOSDIS and oversees an Earth observation program. For EOSDIS the growth in the archive is around 1 PB/yr. For ESA, including missions such as Envisat, Cryosat and AATSR, the data volumes are comparable. Space missions from Japan and India contribute around 0.5 PB/yr and aircraft missions, including LIDAR, probably account for another 0.5 PB/yr. New missions, still in the planning stage, which are expected to launch in the next 5 – 10 years are likely to generate another 1 PB/yr.

If we add all these sources up we arrive at around 4 PB/yr for Earth observation data, depending on the precise definitions (i.e., storage or distribution). For example, EOSDIS distributes 3.62 PB/yr now – made up of 4200 data products, > 400 million product distributions and over 1M users.

#### Oceanographic data

Present ocean sensors tend to generate relatively low volumes of data and typically do so intermittently – during cruises or delivery to shore-side systems. All told, this probably only amounts to tens of GB/day. However, there will be a transformational change with the deployment of real-time ocean observatories.

The Ocean Observatory Initiative or OOI is a long-term, NSF-funded program to provide 25-30 years of sustained ocean measurements to study climate variability, ocean circulation and ecosystem dynamics, air-sea exchange, seafloor processes, and plate-scale geodynamics [25]. The OOI will enable powerful new scientific approaches for exploring the complexities of Earth-ocean-atmosphere interactions, and accelerate progress toward the goal of understanding, predicting, and managing our ocean environment. The Observatory is planned to be a networked infrastructure of science-driven sensor systems to measure the physical, chemical, geological and biological variables in the ocean and seafloor. When complete, the OOI will be one fully integrated system collecting data on coastal, regional and global scales. As a result, the data volumes are expected to increase dramatically and transform ocean-science from being a data-impoverished branch of science to one having an abundance of data.

## Solid Earth Science

Some scientific areas have explosive drivers for the data deluge. This is the situation for astronomy with the ambitious SKA and LSST projects and in particle physics with the LHC accelerator. Other fields are seeing a data deluge but at a lower magnitude. Here we discuss earthquake and polar science.

#### Earthquake Science

Fortunately for society if not science, large earthquakes are infrequent and so the study of earthquakes is data-limited compared to other fields. Major quakes occur all over the world and it is unrealistic to have substantial sensors deployed in most of these regions. Further, the quasi-periodicity of earthquakes implies that historical data is very important and we cannot increase that. Simulations can forecast damage and perhaps the aftershocks of an earthquake but the most important capability – forecasting new quakes is essentially entirely observational. Typically one uses patterns (in time series) to forecast the future with simulations useful to check if a particular pattern informatics approach is valid in an ensemble of simulated earthquakes. Important types of data include:

1. Catalogs of Earthquakes with position and magnitude
2. Geometry of Earthquake faults
3. Global Positioning data (GPS) recording time dependent positions
4. Synthetic Aperture Radar inferograms (InSAR) recording changes in regions over time.

The first two types of data, 1) and 2), are gathered carefully with recording of earthquakes and field analysis.; This is small in size and only growing slowly but of very high value.

For the GPS data there are currently fewer than 10,000 GPS stations recording data at intervals varying between one second and a day. Well-known GPS networks are the Southern California Integrated GPS Network, the Bay Area Regional Deformation Network in Northern California and the PBO Plate Boundary Observatory from UNAVCO [26].

The inSAR data could become voluminous but currently totals some 350 images (each covering around 10,000 km2) and only 2 Terabytes in size. This data comes from uninhabited aerial vehicles (UAVSAR from JPL) or satellites (WInSAR from UNAVCO). The situation could be revolutionized by the approval of the DESDynI-R Mission (Deformation Ecosystem and Dynamics of Ice–Radar) recommended in the Earth Science Decadal Survey [27]. DESDynI would produce around a Terabyte of data per day but the mission has not so far been approved and so is many years away from a possible launch. This data is analyzed (as by QuakeSim [28] for recent earthquakes) to find rates of changes, which are then used in simulations that can lead to better understanding of fault structures and their slip rates.

# Polar Science

Turning to Polar science, a good example here is the work of the CReSIS (Center for Remote Sensing of Ice Sheets) [29] led by Kansas University that is pioneering new radar and UAV’s to be used to study ice-sheets.

Description: Picture2*Figure 1.3 Architecture of PolarGrid data analysis Cyberinfrastructure*

For the project, multiple expeditions fly instruments that collect data including:

1. ice thickness and internal layering from radar and seismics, and Synthetic Aperture Radar (SAR) images of ice-bed interface;
2. bed topography generated from ice thickness and surface elevation);
3. time series of change in surface elevation from airborne and satellite altimeters;
4. time series of surface velocity from repeat-pass satellite images, in situ GPS measurements, and aerial photos;
5. bed characteristics such as temperature, wetness, and sediment from seismics and radar.

The last CReSIS expedition took 80 terabytes of data in 2 months. After traditional processing with FFT’s, radar images are produced along multiple flight lines as illustrated in figure 1.3. Then image processing is needed to identify the top (red) and bottom (green) of an ice-sheet. Initially students performed this but recently it has been automated with an image analysis tool developed by David Crandall at Indiana University. The deployment of UAV’s rather than current Orion and DC-8 conventional aircraft will increase data gathering capability by allowing continuous operation. There are more complex data such as snow deposits showing the annual layers and revealing historical snow deposition.

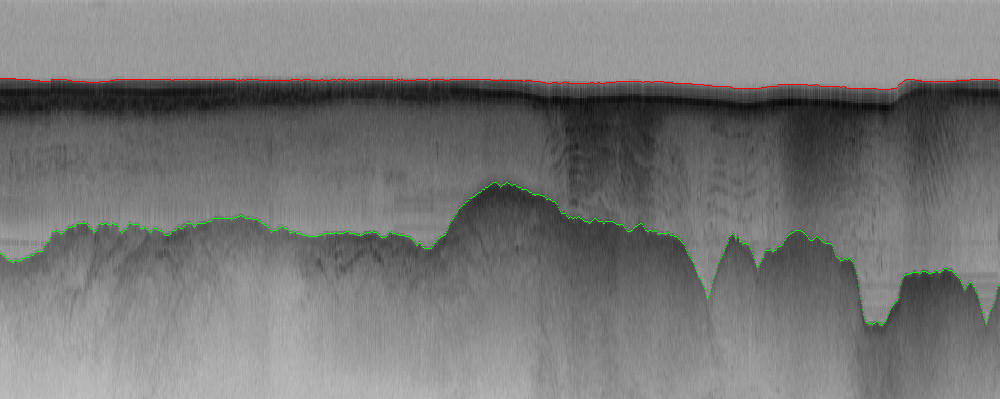


Figure 1.4: Radar imagery from CReSIS

The sea-bed data illustrated in figure 1.4 is fed into simulations that aim to understand the effect of climate change on glaciers. Note that gathering of data is complicated by the paucity of electrical power and poor internet connectivity to the Polar Regions. The use of GPU’s is an interesting technology to deliver possible lower power data processing. The data is gathered on removable disks mounted in a storage array connected to just one or servers with rugged laptops as personal machines.

## Experimentation

### Particle Physics: The Large Hadron Collide (LHC)

The Large Hadron Collider LHC in Geneva is the highest energy particle accelerator ever constructed and operates in a 17 mile tunnel around the CERN Laboratory near Geneva. Two proton beams can be accelerated to energies of up to 7 TeV and collided to produce a specular spray of particles. The goal is to find a key ingredient of the Standard Model of particle physics, the Higg Boson. The LHC is currently operating at half of its design energy and plans to go to the full energy of 7TeV per beam in 2014 [30].

The LHC hosts four major experiments - Atlas, CMS, Alice and LHCb. The first two experiments each record around 100 events per second with each event about 1.5 Megabytes in size. These 100-450 events are selected in real time from the eventual 109 collisions (events) occurring every second at LHC. The experimental detectors contain 150 million sensors that record data 40 million times second (each read out contains over 20 overlapping events). The reduction of a factor of 4 105 in data size is achieved with a multi stage trigger. Having an effective trigger is a major part of design and selection of an experiment. The trigger is based on detecting “unusual events” with signatures of high transverse momentum and interesting particles (leptons and not baryons or mesons) being produced. The multi-stage trigger includes an initial hardware selection (giving a factor of about 400) followed by a software refinement executing on a dedicated cluster, which for CMS has 7000 cores. The software used in this final “higher level trigger” is a stripped down version of the basic analysis software and must reduce the Terabit/second input from the hardware trigger by about another factor of 1000. Alice is a heavy ion experiment to investigate collisions of lead nuclei in the LHC and has larger events and data rates. The LHCb experiment is lower in both respects than Atlas and CMS.

The LHC produces some 15 petabytes of data per year of all varieties and with the exact value depending on duty factor of accelerator (which is reduced simply to cut electricity cost but also due to malfunction of one or more of the many complex systems) and experiments. The raw data produced by experiments is processed on the LHC Computing Grid which has some 200,000 Cores arranged in a three level structure. Tier-0 is CERN itself, Tier 1 are national facilities and Tier 2 are regional systems. For example one LHC experiment (CMS) has 7 Tier-1 and 50 Tier-2 facilities.

The initial data is analyzed in detail to find the parameters of the particles produced in the event and to disentangle the ~20 collisions in each event. This analysis is often iterative as one improves the many calibration constants for the myriad of detector sensors. Detailed summaries of each event or reconstructed data are produced which yields about half the size of the raw data, i.e. ~0.75 Mb, with this process taking an average of around 15 minutes for each event. The experiments also create simple “analysis object data” or AOD that provides a trade-off between event size and complexity of the available information to optimize flexibility and speed for analyses. An AOD (~0.1 mb) is about 5% of size of the raw data. Finally, there are TAGs, which are about 2 kbyte per event, that can be used to select events for a physics analysis that would be done with the larger AOD selection.

This analysis raw data 🡪 reconstructed data 🡪 AOD and TAGS 🡪 Physics is performed on the multi-tier LHC Computing Grid. Note that every event can be analyzed independently so that many events can be processed in parallel with some concentration operations such as those to gather entries in a histogram. This implies that both Grid and Cloud solutions work with this type of data with currently Grids being the only implementation today.

## Photon Sources

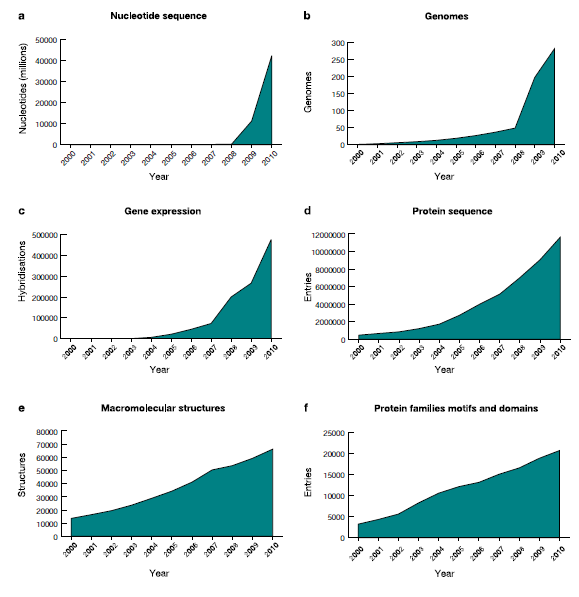
The European Synchroton Radiation Facility (ESRF) in Grenoble uses X-ray radiation in fields as diverse as protein crystallography, earth and materials science as well as certain areas of physics and chemistry [39]. The beamlines are planned to be upgraded to increase the present rate of data production, currently around 1.5 terabytes per day, by two or three orders of magnitude in ten years’ time.

The European X-Ray Free Electron Laser Project (European XFEL) is an international project with 14 participating countries that is located near Hamburg in Germany [40]. Free electron lasers generate high-intensity electromagnetic radiation by accelerating electrons to relativistic speeds in a linear accelerator. By 2015, the European XFEL will produce high-intensity X-ray pulses at intensities much brighter than those produced by conventional light sources. The data rates and data volumes generated by advanced facilities like the European XFEL or the LCLS at Stanford in the USA [41] will exceed those at conventional synchrotron light sources by at least an order of magnitude. Data rates are likely to be of the order of 7 terabytes per hour, depending on the experiment.

These futuristic projections need to be tempered with reality. At present the UK Diamond Light Source is storing approximately 200terabytes in 88m files, increasing to a petayte by 2014 [42]. For these facilities, it is often the number of files which is a challenge rather than the total data volume.

## Biological systems

Within Europe the European Bioinformatics Institute (EBI) is the primary host for bioinformatics data, curating and sharing data from throughout Europe and beyond [43]. It is an academic research institute located in Hinxton near Cambridge (UK) and is part of the [European Molecular Biology Laboratory](http://www.embl.org/" \t "_blank) (EMBL) [44]. The EBI annual report [45] reports on the status of the several databases within the EBI. For example, the European Nucleotide Archive (ENA) had an accumulation rate of more than 500, 000 bases per second. The figures below, taken from the annual report, indicate the growth across other databases hosted at the EBI and the story is much the same.



The increase in physical storage requirements of the Institute from 1996 to 2010 is shown below. The EBI, as part of the ENSEMBL project are now presenting data and services in the Cloud (<http://www.ensembl.org/info/data/amazon_aws.html>) – in this case in the Amazon Cloud.

# 

The explosion in genomics data is being driven by the large-scale automation of the task of gene sequencing. If we look at the implications of one such technology we can get some idea of the transformation that is currently under way. Each Illumina HiSEQ machine generates 108 reads, each roughly 100 Nucleotides long, each day. Each Nucleotide is 2 bits. It takes 100-10000 cores to use the Blast algorithm to compare with data from a central database in one day. Each read is distilled from a coverage of 50-100 times as much data including duplicates.

If we take a unit as a Human Genome with 3x109 Nucleotides or 6x109 bits we can see more graphically the implications of this rate of data production. Each day one Illumina machine sequences 1010 Nucleotides, equivalent to 3.3 Human Genome units per day. If we assume that today there are around 1000 Illumina machines deployed world-wide (500 in USA), each capable of sequencing 3300 Human genomes per day, then this amounts to 2x1013 bits per day corresponding to about 7 Petabits of data per year. (700 Petabits including coverage per year)

Measuring genome of every new born is ~11000 Human genomes per day for USA and 200,000 Human genomes per day for world. Doing on an ongoing basis – say 50 times in lifetime of every human is 5\*10^6 genomes measuredper day for world. This is 30 petabits per day or 10 Exabits per year

It requires power equivalent to 1.5 million present day Illuminas to measure Human genomic data and 1.5\*10^8 to 1.5\*10^10 continuously running present day cores to perform a simple Blast analysis

## Simulations

Weather and Climate

At a September 2008 meeting involving 20 climate modeling groups from around the world, the WCRP's Working Group on Coupled Modelling ([WGCM](http://www.clivar.org/organization/wgcm/wgcm.php)) agreed to promote a new set of coordinated climate model experiments. These experiments comprise the fifth phase of the Coupled Model Intercomparison Project (CMIP5). CMIP5 will notably provide a multi-model context for:

1. assessing the mechanisms responsible for model differences in poorly understood feedbacks associated with the carbon cycle and with clouds,
2. examining climate “predictability” and exploring the ability of models to predict climate on decadal time scales,
3. determining why similarly forced models produce a range of responses.

The fifth Coupled Model Intercomparison Project (CMIP5) will involve the global production and analysis of several petabytes of data. The Program for Climate Model Diagnosis and Intercomparison (PCMDI), with responsibility for archival for CMIP5, has established the global “Earth System Grid Federation” (ESGF) of data producers and data archives to support CMIP5. ESGF will provide a set of globally synchronised views of globally distributed data – including some large cache replicants which will be persisted for (at least) decades. ESGF will stress international networks, as well as the

data archives themselves – but significantly less than would have been the case of a centralised archive. Developing and deploying the ESGF has exploited good will and best efforts, but future developments are likely to require more formalised architecture and management.

ESGF was born out of a number of initiatives to handle diverse, distributed data access for the climate community: In the U.S., the “Earth System Grid” (ESG [xx]), in the UK, “The NERC DataGrid” [xx], and in Germany, the Collaborative Climate Community Data Processing Grid (C3-Grid [xx]) . However, the dominant contribution has been that of the ESG. As a consequence, the ESGF architecture is curently a more mature version of the original ESG, extended and modified by both the code and experiences of the

other partners.

There are five key information classes which underpin the ESGF: the data itself; the “data metadata” which exists within the data files (both described on the CMIP5 website); the “model and experiment metadata” created externally and ingested into the ESGF system [xx]; the “quality metadata” (which describes intrinsic checks on data fidelity rather than the extrinsic scientific quality [xx]); and “federation metadata” (to support user management and system deployment [xx]).

ESGF exploits this information using four major components: data nodes, gateways, federation metadata services (to support authentication and authorisation), and data services to be deployed adjacent (or on) the data nodes.

**Data from Petascale and Exascale Simulations**

In a study [45] of the requirements of applications in exascale computing including High-Energy Physics . Climate , Nuclear Physics, Fusion, Nuclear Energy, Basic Energy Sciences, Biology, and National Security it was reported that exascale applications will generate terabytes of data per second making them one of the largest sources of data. For exascale computing this means a new design for memory, I/O, operating system and software systems. It is likely that exascale simulations will behave as observational science does not in terms of reducing data created before seeking to store any results requiring a new integrated pipeline of data creation through to off-line storage.

This data is produced for two important reasons. One is to provide checkpointing for restart and the second is for visualization and analysis of simulation results. As machines grow in performance, the data produced by simulations naturally scales in size but the associated challenges grow even more as the mean time between failure (MTBF) of total system grows and the compute performance of high end supercomputers tends to grow much faster than disk I/O bandwidth is increasing. There have been several studies of these issues recently [[1-3](file:///C:\Users\tonyhey\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.Outlook\DF61SXGE\Exascale.docx#_ENREF_1)] as part of studies of the next generation exascale systems. The latter will have up to a billion concurrent processes (perhaps arranged as a 1000 threads on each of a million nodes) compared to large simulations today on over 100,000 cores. A new area of study is emerging of processing simulation data in parallel on the nodes of a supercomputer with ADIOS being an interesting approach [[4-6](file:///C:\Users\tonyhey\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.Outlook\DF61SXGE\Exascale.docx#_ENREF_4)]. A study of fusion simulations [[2](file:///C:\Users\tonyhey\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.Outlook\DF61SXGE\Exascale.docx#_ENREF_2)] identified need to output 2 gigabytes of data per simulated time step for each core in the parallel simulation. For “just” a million cores this corresponds to 2 petabytes of data per time step requiring an aggregate I/O rate of 3.5 terabytes/second for a ten minute time step with a simulation of 1 billion cells and 1 trillion particles. An exascale simulation might be 100 times this rate. These data rates are clearly much larger than those associated with observational data although checkpoint data for example can be overwritten and perhaps visualization data will be analyzed in place (by parallel algorithms on same nodes as simulation) and reduced in size before permanent storage.

# Digital libraries

## Data in and out of context

* Open data
* Data Vaults
* Semantics, ontologies etc,
* Energy requirements of data
* And how the Cloud fits in

## Conclusions

There’s a lot of it and it hard to deal with.

**References**

1. The Domesday book: [www.domesdaybook.co.uk/](http://www.domesdaybook.co.uk/)
2. Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, June 2011, www.mckinsey.com/mgi
3. The Domain Name Industry Brief Volume 7, Issue 4, November 2010
4. <http://searchenginewatch.com/article/2073962/New-YouTube-Statistics-48-Hours-of-Video-Uploaded-Per-Minute-3-Billion-Views-Per-Day>
5. How dirty is your data? A Look at the Energy Choices That Power Cloud Computing, Greenpeace International, April 2011
6. Medical Image Utilization: <http://www.acr.org/SecondaryMainMenuCategories/SocioeconomicResearch/utilization.aspx>
7. The National Cancer Institute, Biomedical Imaging Archive Survey: https://wiki.nci.nih.gov/display/CIP/Biomedical+Image+Archives
8. <http://www.ariadne.ac.uk/issue36/tna/>

[8a] <http://energy.aol.com/2011/10/07/keeping-the-lights-on-data-overload-and-its-impact-on-storage-i/>

[8b] <http://www.decc.gov.uk/assets/decc/Consultations/smart-meter-imp-prospectus/1475-smart-metering-imp-response-overview.pdf>

[8c] <http://www.ofgem.gov.uk/e-serve/sm/Documentation/Documents1/Smart%20metering%20-%20Data%20Privacy%20and%20Security.pdf>

1. <http://science.howstuffworks.com/science-vs-myth/everyday-myths/large-hadron-collider3.htm>
2. <http://lcg.web.cern.ch/LCG/>
3. <http://www.docstoc.com/docs/42244651/Overview-of-the-LHC-Triggers-and-Plans-for-LHC>
4. <http://www.ibergrid.eu/2010/pdfs/wednesday/3.%20The%20CMS%20Iberian%20Computing%20Sites%20performance%20in%20the%20advent%20of%20the%20LHC%20era.pdf>
5. <http://iopscience.iop.org/1742-6596/171/1/012101/pdf/1742-6596_171_1_012101.pdf>
6. <https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookComputingModel>
7. Ian Bird, Computing for the Large Hadron Collider, Annual Review of Nuclear and Particle Science, DOI: 10.1146/annurev-nucl-102010-130059
8. <http://www.oceanobservatories.org/>
9. DESDynI <http://desdyni.jpl.nasa.gov/mission/>
10. QuakeSim <http://www.quakesim.org/>
11. UNAVCO and WInSAR <http://www.unavco.org/>
12. Alaska Satellite Facility <http://www.asf.alaska.edu/>
13. UAVSAR <http://uavsar.jpl.nasa.gov/>
14. CReSIS <https://www.cresis.ku.edu/>
15. **Exascale Simulation References**
16. 1. C.S. Chang. *Needs for Extreme Scale DM, Analysis and Visualization in Fusion Particle Code XGC*. [accessed 2011 July 23]; Center for Plasma Edge Simulation CPES at Exascale data management, Analysis, and Visualization, Feb. 22-23, 2011, Houston, TX, Available from: <http://www.olcf.ornl.gov/wp-content/uploads/2011/01/CSChang_Fusion_final.pdf>.
17. 2. *Scientific Grand Challenges: Fusion Energy Sciences and the Role of Computing at the Extreme Scale*. [accessed 2011 July 15]; U. S. Department of Energy, March 18-20, 2009. Washington DC Available from: <http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Fusion_report.pdf>.
18. 3. Jacqueline Chen and John Bell. *Combustion Exascale Co-Design Center*. [accessed 2011 July 15]; 6th International Exascale Software Project Workshop, San Francisco, CA, April 6-7, 201 Available from: <http://www.exascale.org/mediawiki/images/3/38/Talk29-Chen.pdf>.
19. 4. *ADIOS Adaptable IO System* [accessed 2011 July 16]; Available from: <http://adiosapi.org/index.php5?title=Main_Page>.
20. 5. Jay Lofstead, Fang Zheng, Scott Klasky, and Karsten Schwan, *Adaptable, metadata rich IO methods for portable high performance IO*, in *Proceedings of the 2009 IEEE International Symposium on Parallel and Distributed Processing*. 2009, IEEE Computer Society. pages. 1-10. DOI: 10.1109/ipdps.2009.5161052.
21. 6. Y. Tian, S. Klasky, H. Abbasi, J. Lofstead, R. Grout, N. Podhorszki, Q. Liu, Y. Wang, and W. Yu, *EDO: Improving Read Performance for Scientific Applications Through Elastic Data Organization*, in *IEEE Cluster*. September 26-30, 2011. Austin, TX.

45 DOE ASCR 2011 Workshop on Exascale Data Management, Analysis, and Visualization

http://energy.aol.com/2011/10/07/keeping-the-lights-on-data-overload-and-its-impact-on-storage-i/

1. JESS3 are a creative agency that specializes in data visualization [↑](#footnote-ref-1)