# Where does all the data come from?

## Introduction

Data without context is of little or no value. It matters where data has come from and that provenance is something the data must carry with it. Of course digital information is being generated in large quantities each day and depending on the source the information or data come a variety of issues from appropriate semantics to describe it, to integrity and completeness of the data, and all data comes with a cost for keeping it. We tend to think we can and perhaps should keep all digital research outputs to allow future reuse. Some have a vision of a future where data can be searched, found, and federated with new data…. But of course data will be lost and perhaps a more likely future will be that of digital archaeologist piecing together the historic past.

Data born digital comes in many sizes and shapes and from a vast range of sources – in our daily lives we create digital information about ourselves and our lives by shopping with credit cards, using on line systems, social networking, sharing videos, capturing traffic flow, measuring pollution, using security cameras, as well as medical imaging and other related health matters.

Science research creates data from observations, experimentation and simulation and of course research outputs include publications in digital form. In the following chapter we will consider a set of examples and raise some of the issues that we face with respect to these forms. We leave it for later chapters to help the reader understand what the resolution might be to these issues.

## Some Examples

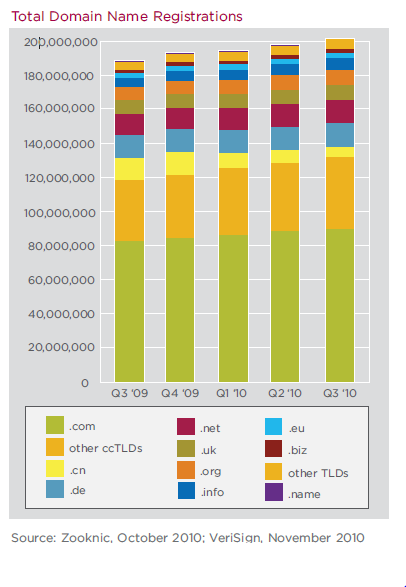
Data in our lives

A recent report by McKinsey Global … [] reports that there are 30 billion pieces of content shared on Facebook every month, that the US Library of Congress had collected 235 TBytes of data by April 2011 and that 15 out of 17 sectors in the US have more data stored per company than 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 provides estimates that globally enterprises store more than 7 Exabytes of new data on disk drives in 2010 and individuals stored more than 6 Exabytes of new data on home and hand held devices.

The enterprise data comprises data generated through interactions with a customer base and data supporting the provision of services through the Internet. McKinsey report that there are over 30 million networked sensors in transportation, industrial, retail, and utilities sections and that this 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 ocean bed examples of data collection.

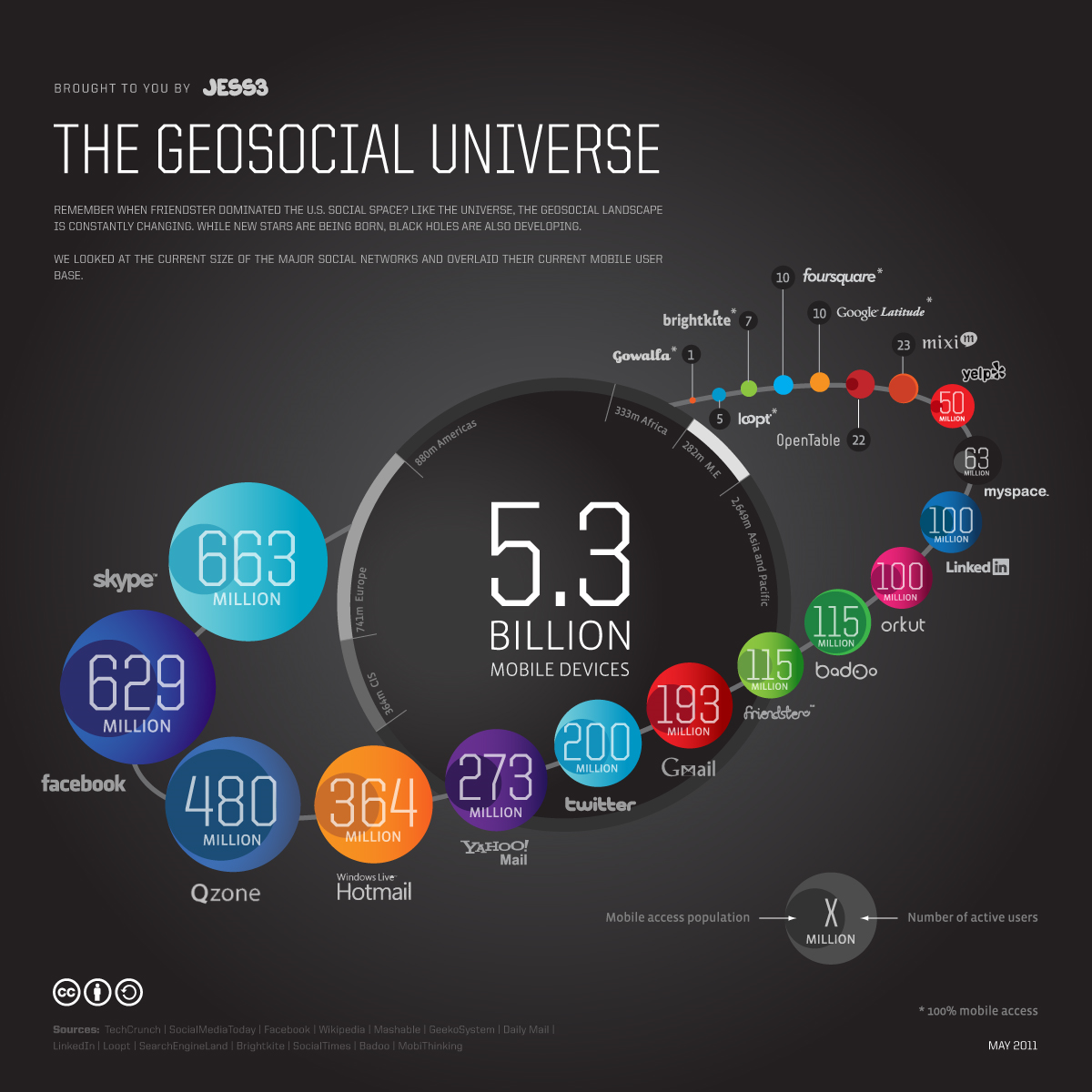
Of course the Internet has changed everything and today our lives are often as much digital as physical. Our collaborations and friendships are as likely to be virtual as not and the management of our lives – banking, house utilities, health, car insurance etc – are increasingly dominated by networked systems and online commerce.

The 2010 Verisign report [ref] indicates that by the end of 2010 there over 202 million domain name registrations and increase of 7% over the previous year. It notes that the largest 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 (?) below shows the breakdown.



Using this information together TNW (The Next Web) [ref] estimate the number of pages on the web to be between 42billion and 121 billion, which is a 21 % increase from 2008. Verisign report that in the last decade the number of Internet users has increased by 500 percent but it is noteable that growth is not homogenous globally, but certain international regions are exploding; for example a decade ago Africa had less than 5 million Internet users but now has more than 100 million. The report indicates that in 2010 less than 40% of Internet users are English speaking.

Social use of the Internet is generating content constantly and in ever increasing amounts. In May 2010 YouTube reported that they had over 2 billion views a day and that 24 hours of video was being uploaded per minute. [Perhaps we should put a figure of GB size here?] . The GreenPeace report “how clean is your data?” reports that 1.2 Zetabytes 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. [need to get permission]



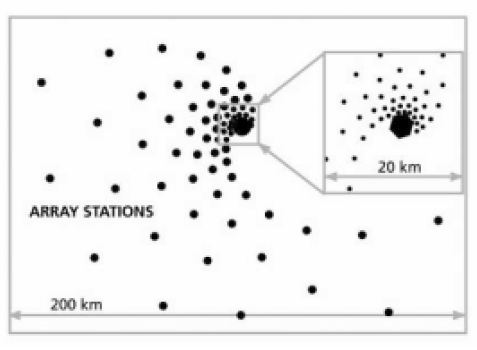
[need piece about medical imaging plus conclusion to social bit]

## Sensors and Observations

### One angle on Astronomy

At the present time hundreds of astronomers, computer scientists and technology engineers are designing the next generation radio telescope - the Square Kilometre Array[[1]](#footnote-1). 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 SKA will likely be located in either Australia or South Africa, in a desert so as to have little or no interference, but is a collaborative effort of over 50 groups in 19 countries.

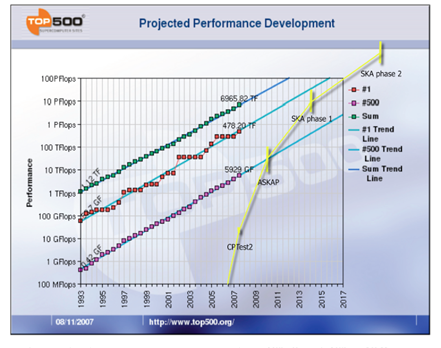
The present design [?,?,?] has a combination of aperture arrays in the core and up to 3000 phased array feeds on dishes and a collecting area of approximately one square kilometre with receptors extending out to a distance of 3000km from the centre of the telescope (figure ?). It will allow a sensitivity of more than 50 times that of existing telescopes, and 10,000 times the survey speed and will provide data to answer fundamental science questions on 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, Pathfinder telescopes that provide experience of design options, and technology capability considerations.



*Figure 1 Possible configuration of SKA receptors and artist’s impression of SKA core(from [?])*

The SKA provides a fabulous information technology challenge with a typical data rate from each dish antenna on the order of 100Gbs-1 aggregating to over 100Tbs-1 [4] and need for Exaflop computation [5] for post-processing. The infrastructure required to support the various science cases will need to range from real-time capability to transport and analyse the data at these high-data rates and 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 and Cloud-like data centres. The communications infrastructure will range from intra-chip and inter-chip with optical fibre to the correlator and on to a high-performance computer, to trans-oceanographic with the latter having data rates of at least 100Gbs-1 over the general network providers. 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 [6] but even this may not take full account of the complexity of the task.

The analysis of the raw data will require exascale computing capability and without this the data will be of no value.

 Cornwell et al [?] estimate the computational requirements in the context of the Top500 and show them to be beyond the scale of projected performance in the timescale required (figure 2).

Alongside the challenges of the computational infrastructure are the related challenges of powering the infrastructure. This includes the power required for the core antennas (~30MW) and the remote stations ~0.5MW) and the various computational components including high-performance computing (~40MW) data transmission (~?MW) and Cloud (or equivalent) provision (~?MW). Energy provision is a major design factor in the delivery of the telescope – plans to mitigate the energy constraints include renewable energy sources (sun) providing the station power and it is easy to see that the SKA could possibly be the largest Green IT project ever to be considered.

### LHC Data Analysis

The Large Hadron Collider LHC has 4 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 10^9 collisions (events) occurring every second at LHC. Those events 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 10^5 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 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. The heavy ion experiment Alice has larger events and data rates while LHCb is lower in both respects than Atlas and CMS.

The LHC produces some 15 petabytes of data per year of all sorts with exact value depending on duty factor of accelerator (which is reduced simply to cut electricity cost but also due to malfunction 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. One produces detailed summaries of each event or reconstructed data which is about half size of raw the data i.e. ~0.75 mb with this process taking an average of around 15 minutes for each event. One also creates 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 5% of size of the raw data. Finally, there are TAGs, 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 multitier LHC Computing Grid. Note that every event can be analyzed independently in parallel with some concentration operations such as those to gather entries in a histogram. This implies that Grid and Cloud solutions work in this example with Grids being the implementation today.

### Solid Earth Science

Some scientific areas have explosive drivers for the data deluge. We have seen this for LHC accelerator, gene sequencing, weather/climate, Astronomy and Exascale visualization. Drivers include both technology improvements and also large increases in deployment due to pervasive need of say personal health or weather data. Other fields see data deluge but at a lower magnitude. Here we discuss earthquake and polar 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 for most of these. 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

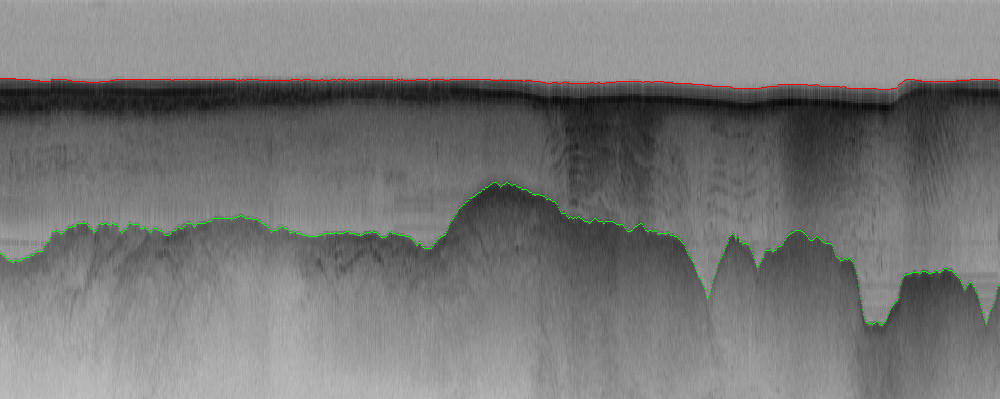
1. Catalogs of Earthquakes with position and magnitude
2. Geometry of Earthquake faults
3. Global Positioning data (GPS) recording time dependent positions
4. Synthetic Aperture Radar inferograms InSAR recording changes in regions over time.

The first two pieces of data a) and b) are gathered carefully with recording of earthquakes and field analysis; this is small in size and growing slowly in size but of very high value. There are currently less than 10,000 GPS stations recording data at an interval 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.

The inSAR data could become voluminous but currently totals some 350 images (each covering around 10,000 km2) and 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. DESDynI would produce around a terabyte of data per day but the mission is not approved and so is many years away from a possible launch. This data is analyzed (as by QuakeSim for recent earthquakes) to find rate of changes, which are then used in simulations that can lead to better understanding of fault structures and their slip rates.

Description: Picture2*Figure 1: Architecture of PolarGrid data analysis Cyberinfrastructure*

Turning to Polar science, a good example here is the work of the CReSIS (Center for Remote Sensing of Ice Sheets) led by Kansas University that is pioneering new radar and UAV’s to be used to study ice-sheets. 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; and (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 2. 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 annual layers revealing historical snow deposition.



*Figure 2: Radar imagery from CReSIS*

The sea-bed data illustrated in figure 2 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. GPU’s are interesting as a possible lower power processing approach. The data is gathered on removable disks mounted in a storage array connected to just one or servers with rugged laptops as personal machines. All software is written in Matlab.

### SmartGrids [tp be added]

## Simulating the world [to be done]

* Weather
* Climate
* bio

## Biological systems

Each Illumina HiSEQ generates 10^8 reads each of ~100 Nucleotides long each day. Each Nucleotide is 2 bits. It takes 100-10000 cores to use Blast to compare with central database in one day (depending on hashing algorithm trading compute performance v. accuracy). Each read is distilled from a coverage of 50-100 times as much data including duplicates.

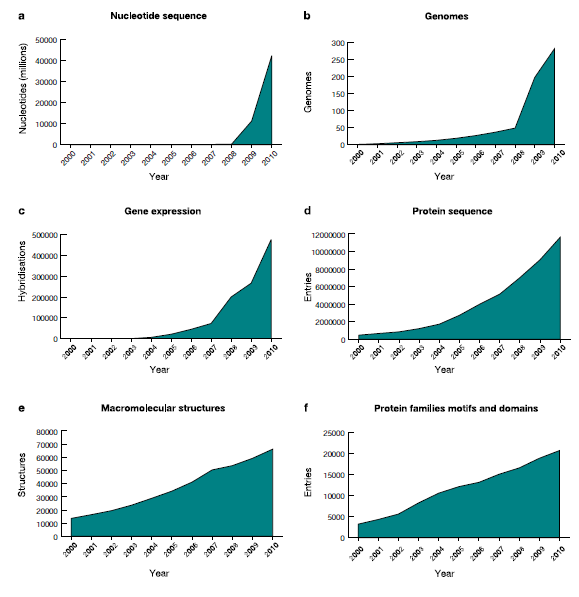
Take a unit as Human Genome with 3\*10^9 Nucleotides or 6\*10^9 bits.

Each day one Illumina does 10^10 Nucleotides, 2\*10^10 bits and 3.3 Human genome units per day. Today there are ~1000 Illuminas (500 in USA) capable of 3300 Human genomes per day; 2\*10^13 bits per day and ~ 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

Within Europe the European Bioinformatics[[2]](#footnote-2) Institute (EBI) is the primary host for bioinformatics data, curating and sharing data from throughout Europe and beyond. 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/) (EMBL). The EBI annual report [ref] 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.

# 

# 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

# How Green is your data?

# Counting Joules

As indicated above the ecosystem of computational resources required to enable the SKA provides a plethora of challenges and opportunities where energy efficiency is concerned.

Moving data takes energy whether the data is moving from L2 cache on chip or within a transatlantic Data Cloud. Effective management of that data communication is a major component of any optimal energy model. There has been a great deal of research on wireless network communications and sensor networks, where devices are most often low-energy devices with battery constraints. Indeed a lot of research has been done in general on low-energy devices including computational algorithms from which we might learn. The existing communications network across the globe relies on hundreds of thousands of switches and routers and these, unlike wireless or mobile infrastructure, do not power down when idle. The network is a complex system of different technologies and it is difficult to predict the energy consumption under different circumstances so sending data 10 times as fast might use interfaces that use 100 times the amount energy or in some cases (e.g. newer optical ones) 1000 times less []. A good overview of network energy costs for realistic configurations can be found in []. There are a number of research efforts considering the issues around Green network communications including INTERNET [] and others can be seen at recent conferences [,]. Of course as energy-aware systems are developed there are still issues of what they optimise – usage or cost? Qureshi et al [] illustrate effective ways in which energy-aware data centres can optimise cost by moving computation to nearby states where electricity costs are less.

Optimizing energy usage in large-scale data centres and Clouds is almost a science in itself. McKinsey and the UpTime Insitute [] indicate that the energy used by data centres in the US is becoming a significant percentage and is likely to overtake airlines in terms of carbon emissions. The report states that the average data centre uses as much power as 25,000 households, but that estimate is probably somewhat out of date as in the last few years Cloud provisioners have built very –large scale data centres across the US []. On a somewhat smaller scale at the Oxford Supercomputing Centre (OSC) we have developed software that intelligently powers down components of the systems at times of under utilisation. The indications are that this will provide significant savings.

The McKinsey report identifies a number of issues around the effectiveness of data centres including siloed organisations and limited transparency that match very well with the findings of the HPC/NA for that community. McKinsey make the recommendation that metrics be defined that are not only measuring the facility but are linked to the applications using it and the processes integral to it. A consortium called the Green Grid is now in place with the aim to develop standards and best practice for data centres. The recommendation regarding metrics is one that we, as a community, should also take on board as we develop exascale technology.

## Conclusions

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

**References**

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. **References**
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/>

1. www.skatelescope.org [↑](#footnote-ref-1)
2. http://www.ebi.ac.uk/ [↑](#footnote-ref-2)