Preliminary OSDC PIRE 2013 Edinburgh Workshop Program

Introduction & Background

The OSDC PIRE program participants who are destined for sites other than Edinburgh will come to this workshop. We will also be inviting local researchers and distinguished external speakers who have kindred interests. The workshop will stimulate research, educate participants and initiate international research relationships.

The workshop is sponsored by the NSF-funded OSDC PIRE project. It will also be supported from various Scottish sources: the School of Informatics (www.ed.ac.uk/schools-departments/informatics), SICSA (www.sicsa.ac.uk/home), CISA (www.cisa.inf.ed.ac.uk) and the Data-Intensive Research Group (research.nesc.ac.uk). It follows on from a successful OSDC PIRE 2012 Workshop, also held in the Informatics Forum, University of Edinburgh. A report on that workshop, including participant feedback is available and has been considered in the preparation of this program.

***All* participants must register** for the event at the following web site so that we can plan computing facilities, catering and meeting room space: osdc-pire-edinburgh2013.eventbrite.com

Overview & Structure

Day	Topics	Evening
Monday	Welcome. Introduction to OSDC software and methods. PICO presentations[1]: self-introductions.	Opportunity to gather at an old Inn (Sheep's Head) and to play traditional bar skittles
Tuesday	Data-driven biological & medical applications. OSDC technical training. PICO *research* presentations by PIRE Fellows and other participants with follow-up discussion time.	Barbeque and reception on fourth floor balcony.
Wednesday	Technology topics of general interest with hands-on opportunities. Data-intensive applications; making them usable. Breakouts discussing (a) the utility of the technologies, and (b) the priorities for technical innovation.	Workshop dinner at Pierre Victoire in Edinburgh
Thursday	Advanced data-intensive technologies Parallel sessions developing background and plans for each host site. Integration of ideas, lessons learned and future road map	Free time
Friday	Social excursion (optional but strongly encouraged)	

[1] *A PICO presentation allows a presenter to make a quick pitch for their topic, two slides and two minutes. They then can present their topic in more detail by a) having a full set of slides on the workshop web site, and b) talking further to interested parties beside a poster, by showing the slides or by showing a demo, in a specified location.*

DAY 1 (Mon, 17-Jun 2013)

Time	*Description*
08:45-09:30	**Registration** This will be on the 4th floor of the Informatics Forum and will be well signed. Coffee and muffins will be served
09:30-11:00	**Session 1:** **Title: Welcome & Introduction, Chair: Malcolm Atkinson,** *U of Edinburgh*
09:30-10:00	**Welcome and "Why are we here?" Speaker: Malcolm Atkinson,** *U of Edinburgh* The digital revolution is bestowing on the whole of mankind a DATA bonanza. It is a major element of the digital revolution – a revolution more fierce than any that has gone before it. This revolution offers a plethora of opportunities and those people, societies and governments that grasp these opportunities will lift their fortunes at the expense of who are less adroit. We need to sharpen our understanding and our skills, we need to develop new mores and the professional integrity to use data well, and we need expertise in spotting and grasping the emerging scientific and business opportunities. This is a path that has only just begun: a journey of a thousand miles begins with a single step. This week we will work hard together to make that first step.
10:00-11:00	**Title: Using the OSDC for Data-Intensive Research, Speaker: Robert Grossman,** *U of Chicago* We give an introduction to the Open Science Data Cloud (OSDC), a science cloud that is operated by the not-for-profit Open Cloud Consortium. The OSDC is a persistent cloud-based infrastructure that allows scientists to manage, analyze, integrate and share medium to large size scientific datasets. The OSDC contains data from a variety of scientific disciplines, from earth science to biology. The OSDC is currently one of the largest cloud-based infrastructures devoted to scientific data.
11:00-11:30	*Coffee break*
11:30-13:00	**Session 2: OSDC PIRE participant self-introduction session** **Chair: Heidi Alvarez,** *FIU*
	Participants in the workshop, including organizers and speakers, will introduce themselves and provide a record about themselves for consultation during and after the workshop. You were invited to provide two slides. You can also provide a title slide with your name and home institution on it; we'll make this title slide if you didn't. The other two slides should contain what you think is important that people at the OSDC workshop know about you. The three slides will be spliced into an auto-advancing presentation: 1 minute for the title slide, during which time the previous speaker sits down and the next takes their place on the podium. Your next two slides will auto advance after a minute (no animations!). You can speak to them or say related things. This is called a PICO presentation; see below. BE YE WARNED: if you don't provide slides (PowerPoint, Keynote or PDF – we'll do the conversion) we will provide two blank slides for your two minutes. The order will be random but pre-published. You have an important role to play during the first day of the Edinburgh Workshop. In order to get to know each other we would like you to prepare a PICO presentation as an interesting way to present yourself and the area(s) of your research. Below are the links with details for PICO presentations also known as introductory "2-minutes-madness". Presenting Interactive COntent, or PICO, sessions highlight the essence of a particular research area – just enough to get excited about a topic without being overloaded with information. PICO sessions combine the best of oral and poster presentations. Every PICO author presents 2 slides in a "2 minutes madness" and afterwards, all attendees have enough time to watch the presentation again (we will post on the OSDC PIRE Facebook Fan Page & Workshop website) and hold discussions with both the author and other attendees. See this short video. https://www.youtube.com/watch?feature=player_embedded&v=A36FZZiZYVE Find more details here http://www.egu2013.eu/guidelines/author_guidelines_pico.html Additionally, we would like to post your picture and a short Bio (one paragraph) on the OSDC PIRE Facebook Fan Page. After the PICO presentations, we will have a general discussion about the workshop and what you want to get from it.
13:00-14:00	*Lunch*

14:00-15:00	**Session 3:** **Title: Using the Open Science Data Cloud for Data-Intensive Research** (continued) **Speaker: Robert Grossman,** *U of Chicago*
15:00-15:30	**Title: Tutorial and Hands on OSDC exercise** **Speaker: Allison Heath,** *U of Chicago* This is an introduction to using the OSDC, using OSDC-Sullivan. Please make sure to bring your laptops so you can follow along. **IMPORTANT: If you do not have an account on OSDC-Sullivan, please apply at least a week before the start of the workshop at** https://www.opensciencedatacloud.org/apply/. Your email address is used for providing access to the OSDC. We support most research institutions, so please provide your primary institutional email on the application.
15:30-16:00	*Coffee break*
16:00-17:00	**Title: Tutorial and Hands on OSDC exercises** (continued), **Speaker: Allison Heath,** *U of Chicago*
17:00-17:15	**Title: Wrap up & Review, Facilitators:** **Heidi Alvarez, Robert Grossman and Malcolm Atkinson**
18:00- Late	*Refreshments & Bar Skittles at the Sheep Heid, Duddingston www.thesheepheidedinburgh.co.uk*

DAY 2 (Tues, 18-Jun 2013)

09:00-10:30	**Session 4: U Chicago Applications & Infrastructure 10 -20 Minute Talks** **Chair: Robert Grossman,** *U of Chicago*
9:00-9:20	**Title:** **Bionimbus - Managing, Analyzing and Sharing Large Genomic Datasets** **Speaker: Allison Heath,** *U of Chicago* Bionimbus is a petabyte-scale community cloud for managing, analyzing and sharing large genomics datasets that is operated by the not-for-profit Open Cloud Consortium. It contains public datasets, such as the 1000 Genomes dataset. We recently updated Bionimbus so that researchers can analyze data from controlled datasets, such as The Cancer Genome Atlas (TCGA) in a secure and compliant fashion.
9:20-9:40	**Title:** **Tukey: The OSDC User Interface** **Speaker: Matt Greenway,** *U of Chicago* Tukey provides the web front end for OSDC users where they can provision and manage virtual machines, manage credentials and view usage. Tukey is written in python using the Django framework and is based off of Horizon, the OpenStack dashboard. Tukey adds a number of features that do not exist in Horizon or other cloud dashboards, such as federated authentication using Shibboleth and OpenID and support for both Eucalyptus and OpenStack clouds.
9:40-10:00	**Title:** **Yates: The OSDC Automation Infrastructure** **Speaker: Rafael Suarez,** *U of Chicago* Yates is based on Chef and enables us to bring up a rack of equipment configured with OpenStack into the OSDC in less than an hour. Yates is an example of the automation required to operate large scale science clouds efficiently.
10:00-10:20	**Title:** **The Namibia Flood Detection Dashboard** **Speaker: Zac Fleming,** *U of Oklahoma* The Namibia Flood Detection Dashboard integrates data from a number of satellites and other data sources in an interactive Open Geospatial Consortium (OGC) compliant map. The dashboard is used to provide early warnings of areas that are likely to flood. In this talk, we describe the dashboard and how it is used.
10:20-10:30	**Questions & Answers about the above presentations**
10:30-11:00	*Coffee break*
11:00-11:30	**Title:** **An Overview of Matsu: An Open Standards-Based Cloud Infrastructure of Biological Data** **Speaker: Robert Grossman,** *U of Chicago* The Matsu Project has developed an open source cloud-based infrastructure to process, analyze, and reanalyze large collections of hyperspectral satellite image data using OpenStack, Hadoop, MapReduce and related technologies. The Matsu Project includes a framework for running a collection of distributed analytics over all the satellite images. The results of the analytics are accessible through an Open Geospatial Compliant (OGC)-compliant Web Map Service (WMS). Matsu is a project of the Open Cloud Consortium. Matsu is currently processing the data produced each day by NASA's EO-1 satellite.
11:30-12:00	**Title:** **UDR: A utility for synchronizing large remote data sets** **Speaker: Allison Heath,** *U of Chicago.* The rsync utility is a wonderfully useful tool for keeping two datasets synchronized, but it was never designed to keep two large datasets synchronized when they are separated by a long distance. Over the past couple of years, we have developed a utility called UDR that integrates rsync with the high performance network protocol UDT. UDT is a reliable UDP-based protocol that was designed to move large datasets over wide area, high performance networks. UDT is open source and has been used as the basis for over six commercial products.
12:00-12:30	**Title:** **MoSGrid – Molecular simulations in a distributed environment** **Speaker: Sandra Gesing,** *U of Edinburgh (en route from U of Tübingen, Germany to Notre Dame, Indiana, USA)* Molecular simulations are indispensable methods in areas like material science, structural biology, and drug design. These methods address data-intensive and compute-intensive problems, which demand high-performance computing to allow data analysis in an acceptable time. The project MoSGrid (Molecular Simulation Grid) offers a workflow-enabled grid portal allowing access to molecular simulation tools on distributed resources in an intuitive manner. Users are able to exchange workflows and data via repositories and, thus, to exchange knowledge about the specific application domain. The talk will give an overview on the current MoSGrid portal, the underlying infrastructure, and an outlook on the next steps.

12:30-13:00	**Discussion, Chair: Malcolm Atkinson,** *U of Edinburgh*
13:00-14:00	*Lunch*
14:00-15:00	**Session Hands-on Exercises from Session 3 conclude** **Title:** Technical Interactive Session – Walk through an OSDC usage exercise**, Speaker: Allison Heath,** *U of Chicago*
15:00-15:30	***Early*** *Coffee break*
15:30-19:30	**Session 5: The Research Bazaar,** **Organizer: Malcolm Atkinson,** *U of Edinburgh*
	The goal is to stimulate wide-ranging research discussion around topics of interest to the OSDC PIRE Fellows Those running the workshop or giving talks longer than 20 minutes already, cannot initiate a discussion – they have already had their chance. Every OSDC PIRE Fellow and their host counterparts are strongly encouraged to make a pitch for research they have done, research they are doing or research they would like to do. This may be research in their own institution or in the host institution they are about to visit. They should make a strong pitch to get others interested in follow up discussions. The session will be in two parts: Part 1: PICO pitches to raise interest – fast and furious to leave plenty of time for Part 2: Parallel and self-organized discussions with gentle shepherding. ***You should send us your PICO slides for this session by mid-day (BST) Monday 17 June***, so there is time to splice them together and to put them on the workshop web site. You should have a title slide that includes your name and home institution and two following slides without animation, for your PICO talk. After these you can have other slides which will go on the web site, but which you will not be able to use during your PICO pitch. Again the slides will auto advance. The order will correspond to when we receive your slides; the early birds get the first talks. When these have been presented, each PICO presenter should be prepared to lead a follow up discussion (possibly more than once) using one of the following three methods: 1. Showing your full set of slides and talking about them, 2. Standing by a poster and talking about it (we'll try to get these posters on the workshop web site), or 3. Showing a demonstration of your research results. Please tell us which of these you want, so that we can set up facilities for them. Depending on demand, you'll get from 20 to 30 minutes for a follow-up discussion, with an option of repeat performances and informal follow up. We will schedule a round of these and allocate places to a subset of the pitches depending on the overlaps in interest indicated by the audience. Then there will be a short hiatus while people change over, and we'll run another set of parallel discussions, and so on. This will be repeated until everyone is satisfied or exhausted.
17:30-19:30	***Reception & Barbeque*** Available on the balcony but if you want to continue your discussions while you avail yourself of tempting morsels and imbibe a suitable elixir please do so!

DAY 3 (Wed, 19-Jun 2013)

09:00-13:00	**Session 6: Data-Intensive Streaming Methods** **Chair: Malcolm Atkinson,** *U of Edinburgh*
09:00-09:15	**Title:** **Motivation and Architecture of Data-Streaming Systems** **Speaker: Malcolm Atkinson,** *U of Edinburgh* Processing continuous data streams is essential when data sources are continuously emitting data. These are also efficacious with large volumes of data, because they exploit caching well, they pipeline work with minimal I/O, they allow users to steer active analyses and the amortized set-up costs over more data. We introduce DISPEL, our experimental language to design the appropriate universe of discourse, abstractions and principles for data-streaming applications.
09:15-10:30	**Title:** **A practical introduction to DISPEL** **Speakers: Paul Martin (assisted by: Michelle Galea, Amy Krause, Alessandro Spinuso and Iraklis Klampanos),** *U of Edinburgh* The DISPEL semantics and enactment system is introduced. Participants will be supported in trying some very simple examples. This experience will equip them to explore further DISPEL tutorial material and to discuss DISPEL applications. It is intended that this will equip participants with an improved conceptual vocabulary for discussing data-intensive applications and systems.
10:30-11:00	*Coffee break*
11:00-13:00	**Session 7: Collaborative and distributed e-Science** **Chair: Heidi Alvarez**
11:00-11:30	**Title:** **Data-Intensive Seismology** **Speaker: Iraklis Klampanos,** *U of Edinburgh* Typical of the digital revolution, seismology is changing rapidly: the number of deployed seismometers grows rapidly, their performance and connectivity improve and seismologists employ data from other sources, such as LIDAR, satellite images and GPS. But their computational behavior is also changing: they used to just focus on earthquakes, today the use all of the continuous waveform. Just what does this mean for the science and the computational infrastructure to support it? The VERCE project, www.verce.eu, is pioneering this approach and will provide examples during the talk.
11:30-11:45	**Title:** **Multi-workflow Systems and Editors** **Speaker: Sandra Gesing,** *U of Edinburgh (en route from U of Tübingen, Germany to Notre Dame, Indiana, USA)* Researchers often need to use workflows that have been developed by other experts in their field to handle specific parts of their work. Sooner or later they find that they want to use workflows from multiple sources that are written in different languages. Enacting multi-lingual workflows (or meta workflows) has been pioneered in a group of European projects. The next step is to be able to change them when they don't do exactly what you want. But that is not easy if you need to learn a different editor for each workflow language. The talk will conclude with the status of research underway to address this problem.
11:45-12:00	**Title:** **The EFFORT Science Gateway** **Speaker: Rosa Filgueira,** *U of Edinburgh* Rock physicists and volcanologist could gain from sharing experimental data and computational models; their computational practices are similar to those of seismologists. The talk will report on the experience of building and running a science gateway in the EFFORT project with the intention of opening up opportunities for sharing experimental data in a way that is integrated with multiple hypothesis testing. A demo of the current status of the gateway will be displayed at the end of the talk.
12:00-12:30	**Title:** **Sustainable Smart e-Infrastructure**, **Speaker: Paola Grosso,** *University of Amsterdam* With clouds we can compute cheaply and efficiently; with dynamic and programmable networks we can transport the data to the clouds. But what are the sustainability effects of sending data to remote sites? My recent research tries to determine under which conditions moving the bits to locations powered with green energy results in smaller carbon footprint than purchasing green energy for the local site. The novel insight in this work is the quantification of the contribution given by the underlying transport network in this decision.

12:30-12:45	**Title: FAST, Speaker: Christine Harvey and Rosa Filgueira,** *XXX and University of Edinburgh* This talk reports on work undertaken by Christine as one of last year's OSDC PIRE Fellows and on subsequent progress.
12:45-13:00	**Title: Reflections on the achievements of PIRE Fellows, Speaker: Heidi Alvarez,** *Florida International University* To be confirmed.
13:00-14:00	*Lunch*
14:00-15:30	**Session 8: Data-Intensive Science: Principles, Examples & Questions** **Chair: Robert Grossman,** *U of Chicago*
14:00-14:30	**Title: A Principled approach to Provenance** **Speaker: James Cheney,** *U of Edinburgh* Provenance refers to the information that denotes the origin, history, or derivation of an artifact. We use the term in computing to denote a description of the origin and life story of raw or processed data, or even of software or sensors that produced it. Collecting, managing and exploiting provenance information is considered vital for understanding the context and value of data, or for replaying or reproducing computations. While these goals are widely shared among both domain scientists and computer scientists, understanding of how to achieve them systematically and efficiently has proven elusive. The talk will provide a high-level overview of the state of the art of provenance for databases, structured scientific computations (e.g. workflows), and for data on the Web.
14:30-14:50	**Title: Using Big Data to build decision support tools in Agriculture, Speaker: Karen Langona,** *U of Sao Paulo* Information Technology can assist producers to make better decisions by providing them with data and tools that enhance decision-making process, consequently allowing better and sustainable management of the natural resources. The talk will present an agricultural decision-making web-based support system, which uses big data technologies to calculate relevant metrics based on large public climate-related databases. This solution can help producers and/or policymakers to get results from large data sets without the need to either invest on infrastructure or learn complex technologies. The big data framework and the distributed processing features are embedded into the solution in a way that is transparent to the user.
14:50-15:10	**Title: Case Studies in Running Many Simulations on Many Clusters, Clouds and Supercomputers** **Speaker: Shantenu Jha,** *Rutgers University* There are several important science and engineering problems that require the coordinated execution of multiple simulations at scale. Some common scenarios include but are not limited to, "an ensemble of tasks", "loosely-coupled simulations of tightly-coupled simulations" or "multi-component multi-physics simulations". We will discuss representative scalable examples from each of the categories.
15:10-15:30	**Briefing for breakout discussions** **Title: What are the most powerful data-intensive methods and how should we use them?** **Chair: Robert Grossman,** *U of Chicago*
15:30-16:00	*Coffee break*
16:00-17:30	**Session 9: Break-out Group Discussions on the OSDC & Data-Intensive Research Challenges** **Chair: Bob Grossman**
15:45-17:00	Parallel discussions in groups of 5 to 7. Each group should select someone to chair the discussion and two other people to act a rapporteurs. One of the rapporteurs will present the results from the group's discussion in the plenary wrap-up at 17:00. The other will provide a textual summary for the workshop's web site.
17:00-17:30	**Groups report back**: and we discover the level of consensus. **Chair: Robert Grossman,** *U of Chicago*
	Workshop Dinner at Pierre Victoire www.pierrevictoirerestaurant.co.uk

DAY 4 (Thurs, 20-Jun 2013)

Each site that will host PIRE visiting researchers should have a parallel track from morning coffee to afternoon coffee to run a detailed orientation program matching their visitors and projects. ***Only the Edinburgh one is shown at present****. Sites may choose to use part of their time, and participants can then join in other parallel sessions.*

09:00-10:30	**Session 10: Practical technologies for advancing the state of the art** **Chair: Malcolm Atkinson,** *University of Edinburgh*
09:00-09:30	**Title:** Mastering Complex Internet Infrastructure to support Science **Speaker: Cees de Laat,** *University of Amsterdam* In his address, Cees de Laat will talk about the use of state-of-the-art networking in support of e-Science and the great benefit of networking and e-Infrastructure to researchers in many disciplines. With experience from projects in high-energy physics, radio-astronomy, dike engineering, medical research, and more, Cees de Laat will show us how networking innovations enable research collaborations on a new scale with novel capabilities.
09:30-10:00	**Title:** Data-Intensive Research in ITRI/AIST, **Speaker: Isao Kojima,** *AIST* In this talk, several data intensive research topics in AIST are presented. One of them is the GEO Grid project that we have been doing for nearly 10 years. Other topics include: Linked Open Data and Bioinformatics data processing. Several multimedia applications are also shown.
10:00-10:30	**Title:** A Brief Introduction to SAGA and Bigjob, **Speaker: Shantenu Jha**, *Rutgers University* The tools and capabilities to support coordinated multiple simulations are limited. However, a promising way to overcome this common limitation is the use of a Pilot-Job, which can be defined as a container or placeholder job to provide multi-level scheduling via an application-level scheduling overlay over the system scheduler. We discuss both the theory and practice of Pilot-Jobs. Specifically, we present "BigJob" as a SAGA-based extensible, interoperable and scalable Pilot-Job implementation. We then discuss several science problems that have/are using BigJob to execute multiple simulations at unprecedented scales on a range of supercomputers and distributed supercomputing infrastructure such as XSEDE.
10:30-11:00	*Coffee break*
11:00-15:30	**Session 11: Parallel Sessions**
11:00-16:00	**Session 11A: Research in AIST** **Chair: Isao Kojima**
11:00-12:30	Orientation to research at AIST.
12:30-14:00	*Extended lunch with informal discussions*
14:00-15:30	Join Edinburgh track
11:00-16:00	**Session 11B: Research in Amsterdam and Sao Paolo** **Chairs: Cees de Laat and Karen Langona**
11:00-11:45	Whatever sessions Cees & Karen agree on until 15:30
11:45-12:00	
11:45-12:00	

12:15-12:30	
12:30-14:00	Extended lunch with informal discussions
14:00-15:30	
15:00-15:45	

11:00-16:00	**Session 11C: Research in Edinburgh** **Chair: Malcolm Atkinson,** *U of Edinburgh*
11:00-11:45	**Edinburgh: A Hands-on Introduction to SAGA and BigJob,** **Speaker: Shantenu Jha,** *Rutgers University* The RADICAL team at Rutgers provide scalable implementation of SAGA and BigJob in Python. In this exercise-based tutorial, we will provide a hands on introduction to SAGA-python and a SAGA-based Bigjob. At the end of this tutorial, students will be able to use SAGA to submit remote jobs to a variety of heterogeneous resources, and will be able to use the SAGA-based Pilot-Job (BigJob) to execute a "bag- of tasks" and "loosely-coupled tasks" , amongst other simple execution scenarios.
11:45-12:00	**Provenance for Seismology, Speaker: Alessandro Spinuso,** *KNMI & ORFEUS* The requirements and uses of provenance in VERCE will be described. The technical and scientific challenges of meeting those requirements will be discussed with an analysis of initial experience.
11:45-12:00	**Supporting Collaborative Scientific Workflow Development: The VERCE Information Registry,** **Speaker: Iraklis Klampanos,** *University of Edinburgh* Large-scale distributed workflow systems for science are nowadays expected to operate in a consistent, predictable way as well as to promote collaboration between researchers or within groups in a unified way. In this talk we will discuss the VERCE Information Registry, which is designed to provide a consistent view of the VERCE ecosystem for seismology along with related architectural requirements, assumptions and interactions with other components.
12:15-12:30	**Designing Python Libraries for Rock Physics,** **Speaker: Rosa Filgueira,** *University of Edinburgh* Well-crafted libraries are key to gaining adoption of technologies by the innovating computationally adept scientists. What form and content should a rock-physics Python have? Inspiration may be taken from NumPy, SciPy & ObsPy.
12:30-14:00	*Extended lunch with informal discussions*
14:00-15:00	**Introduction to iRODS handling distributed scientific data, Speaker: TBA**, *EUDAT consortium* iRODS is being used for services that are expected to be common across research infrastructures. The talk will discuss the experiences and challenges of scaling up to a Europe-wide multi-disciplinary scale.
15:00-15:45	**Data Integration and Analysis for Systems Medicine, Ian Overton**, *MRC Human Genetics Unit, Western General Hospital, Edinburgh* Biological processes are organized and controlled by complex interactions between many individual components, and so inherently involve intricate networks. The properties of these networks underlie virtually every aspect of cell function. We apply data-driven approaches to understand these molecular circuits towards better and more effective medicine. Available data is distilled into genome-scale models tailored for particular biological domains, such as Epithelial to Mesenchymal Transition – which is a key interest relevant to metastasis and fibrosis. Our work involves machine-learning and information-theoretic approaches, including development of methods and tools for data analysis as well as to inform clinical decision-making.

15:30-16:00	*Coffee break*
16:00- 17:30	**Session: 12 What have we learnt & where next?** **Chair: Heidi Alvarez, Malcolm Atkinson & Bob Grossman**
16:00-17:00	What have we understood this week about data and how to make the best use of the data bonanza? What do we need to do & understand to make data-use easy and effective? How should we do that?

17:00-17:30	**Wrap-up & valediction**

DAY 5 (Fri, 21-Jun 2013)

Open Now	*Note: There is an option on the eventbrite page for people to register to go on the excursion.* ***We need to know numbers as we need to pay in advance!***
09:30-10:00	**Waverley train station – Exact time and meeting place where we will hand out tickets to be updated**
10:00-11:00	**Sea Bird Centre in North Berwick,** www.seabird.org
11:00-11:30	
11:30-13:00	**Trip out on the catamaran to see Bass Rock Gannet colony and Tantallon Castle**, www.historic-scotland.gov.uk
13:00-14:00	*Use this opportunity to explore facilities and eat in North Berwick*
14:00-17:00	You can explore North Berwick and surrounding beaches, walk along the John Muir trail, or go overland to visit Tantallon Castle. **Choose when you return to Edinburgh on your own or in small groups - flexible**
16:00-16:30	

About the OSDC

About the OSDC. The Open Science Data Cloud or OSDC (www.opensciencedatacloud.org) is a petabyte-scale science cloud managed and operated by the Open Cloud Consortium (OCC) that has been in operation for approximately three years. The OCC is a not-for-profit that develops and operates cloud computing infrastructure for the research community.

The OSDC allows scientists to manage, analyze, share and archive their datasets, even if they are large. Datasets can be downloaded from the OSDC by anyone. Small amounts of computing infrastructure are available without cost so that any researcher can compute over the data managed by the OSDC. Larger amounts of computing infrastructure are made available to researchers at cost. In addition, larger amounts of computing resources are also made available to research projects through a selection process so that interested projects can use the OSDC to manage and analyze their data.

The OSDC is not only designed to provide a long term persistent home for scientific data, but also to provide a platform or “instrument” for data intensive science so that new types of data intensive algorithms can be developed, tested, and used over large amounts of heterogeneous scientific data.

The OSDC is a distributed facility that spans four data centers connected by 10G networks. Two data centers are in Chicago, one is at the Livermore Valley Open Campus (LVOC), and one is at the AMPATH facility in Miami.

The OSDC is based largely on third party open source software, including OpenStack, Eucalyptus, Hadoop and GlusterFS. The OSDC has developed an open source portal that provides users with a single point of access and an open source middleware layer called Tukey that integrates the various OSDC services.

One of the main reasons for the workshop is bring together a critical mass of users that can discuss how to extend the OSDC. Questions of interest include:

Question 1. How can we build a plugin structure so that Tukey can be extended by other users and by other communities?

Question 2. How can we add partner sites at other locations that extend the OSDC?

Question 3. What data can we add to facilitate data intensive cross-disciplinary discoveries?

Question 4. What tools and applications can we add to facilitate data intensive cross-disciplinary discoveries?

Question 5. How can we better integrate digital IDs and file sharing services into the OSDC?

Question 6. What are some grand challenge questions we can pose that leverage the OSDC?