
**2015 ASTRONOMICAL DATA ANALYSIS SYSTEMS**

**AND SOFTWARE CONFERENCE**

**25th – 29th October 2015**

**Rydges World Square, Pitt Street, Sydney, NSW, Australia**

# ORAL ABSTRACT BOOKLET

## SESSION 1: KNOWLEDGE DISCOVERY AND DATA MANAGEMENT TOOLS FOR ASTRONOMICAL BIG DATA

**O1.1 Hugh Durrant-Whyte**
University of Sydney

*Data, Knowledge and Discovery: Machine Learning meets Natural Science*

Increasingly it is data, vast amounts of data, that drives scientific discovery. At the heart of this so-called “fourth paradigm of science” is the rapid development of large scale statistical data fusion and machine learning methods. While these developments in “big data” methods are largely driven by commercial applications such as internet search or customer modelling, the opportunity for applying these to scientific discovery is huge. This talk will describe a number of applied machine learning projects addressing real-world inference problems in physical, life and social science areas. In particular, I will describe a major Science and Industry Endowment Fund (SIEF) project, in collaboration with the NICTA and Macquarie University, looking to apply machine learning techniques to discovery in the natural sciences. This talk will look at the key methods in machine learning that are being applied to the discovery process, especially in areas like geology, ecology and biological discovery.

**O1.2 Alexandra Aloisi**
Space Telescope Science Institute

*Maximising the Science in the Era of Data Driven Astronomy*

In the era of large astronomical data, data-driven multi-wavelength science will play an increasing role in Astronomy over the next decade. A range of new missions, facilities, and surveys including JWST, PanSTARRS, WFIRST/AFTA, LSST, ALMA, SKA, etc. will accumulate peta-bytes of data. The Space Telescope Science Institute and its NASA archive, the Mikulski Archive for Space Telescopes (MAST), will continue to play a key role in this arena. In this talk I will review our archive strategic roadmap over the next five years and the new scientific investigations that this will enable. This includes the deployment of a scalable architecture for easy multi-mission operations, the unification of all the archival services under the MAST portal, the acquisition of new data collections, the production of new science-ready data holdings, the partnerships with other Archives for exchange of data and definition of new interoperability standards, the creation of new tools for data discovery, data mining, and data analysis, and the enabling of new on-line collaborative resources (e.g., virtual machines and science cloud). We will maximize the scientific return of the space Astrophysics programs by providing the Astronomical community with a peta-scale archival collection of data and a powerful open-science environment that will enable high-impact investigations in every area of Astrophysics from the far ultraviolet to the infrared.

---

**O1.3 Lisa Storrie-Lombardi**
Caltech

*Observatory Archives in the Era of Big Data: Perspectives from the Spitzer Mission*

Community Observatory archives, supporting disparate data sets, have traditionally been fundamentally different than homogeneous survey archives. Technological advances in the past decade, particularly with respect to the ease of connectivity, are blurring the lines between observatory archives and large sky surveys. Archives are no longer monolithic repositories of data but instead are portals for data and services as part of the broader scientific landscape. If starting today we would approach the Spitzer archive design with a very different mindset than we did fifteen years ago. We discuss here (1) design lessons learned and the evolution of the archive, (2) the benefits of having the Spitzer archive as a component of IRSA, (3) the value of serving enhanced data sets that are returned to the archive by science teams, and (4) the benefits of connectivity to large survey archives, now and in the future.

---

**O1.4 William O'Mullane**
European Space Astronomy Centre, ESA

*Bringing the Computing to the Data*

It has become clear in the recent years that several data centres are working on bringing users tasks into the data centre instead of executing queries and returning the results. Recently I have seen that several centres, including ESAC, are looking at Docker and Python for this task i.e. in a more or less controlled manner allowing users to deliver Python code to run in a container and access resources within the data centre. If this talk is accepted I would use the ADASS mailing list (or a subset) to survey efforts in different data centres on this topic and present the current thinking in the ADASS session. We have discussed the usefulness of survey talks before but have not had any yet! There is no protocol in IVOA for this yet but perhaps there should be - may be covered by Berriman/Arviset's talk.

---

## SESSION II: ALGORITHMS FOR ASTRONOMICAL DATA REDUCTION

**O2.1 Tamas Budavari**
John Hopkins University

*Streaming Algorithms for Optimal Combination of Images and Catalogs*

Modern astronomy is increasingly relying on large surveys, whose dedicated telescopes tenaciously observe the sky every night. The stream of data is often just collected to be analyzed in batches before each release of a particular project. Processing such large amounts of data is not only inefficient computationally it also introduces a significant delay before the measurements become available for scientific use. We will discuss algorithms that can process data as soon as they become available and provide incrementally improving results over time. In particular we will focus on two problems: (1) Repeated exposures of the sky are usually convolved to the worst acceptable quality before coadding for high signal-to-noise. Instead one can use (blind) image deconvolution to retain high resolution information, while still gaining signal-to-noise ratio. (2) Catalogs (and lightcurves) are traditionally extracted in apertures obtained from deep coadds after all the exposures are taken. Alternatively incremental aggregation and probabilistic filtering of intermediate catalogs could provide immediate access to faint sources during the life of a survey.

---

**O2.2 Guido Cupani**
INAF - Osservatorio Astronomico di Trieste

*Data Analysis for Precision Spectroscopy: the ESPRESSO Case*

Astronomical Spectroscopy is rapidly evolving into a precision science, with several observational projects increasingly relying on long-term instrumental stability and centimeter-per-second accuracy in wavelength calibration. These requirements strongly call for integrated software tools to manage not only the reduction of data, but also the scientific analysis. The ultra-stable, high-resolution echelle spectrograph ESPRESSO, currently under integration for the ESO VLT (first light: 2017) is the first instrument of its kind to include a dedicated Data Analysis Software (DAS) among its deliverables, to process both stellar and quasar spectra. The DAS will extract physical information from the reduced data on the fly (e.g. stellar radial velocities, or characterisation of the absorption systems along the sightline to quasars) and will allow interaction through a configurable graphical user interface. In this oral presentation I will showcase the DAS features and its development status one month away from the first public release. A particular attention will be devoted to the algorithms developed for quasar spectral analysis (Voigt-profile line fitting and continuum determination). Some short videos showing the software in action will be displayed.

---

**O2.3 Mohammad Akhlaghi**
Tohoku University Astronomical Insitute

*NoiseBased detection and segmentation of nebulous signal*
*Co-author: Professor Takashi Ichikawa*

Because of the rich dynamic history of internal and external processes, galaxies display a very diverse variety of shapes or morphologies. Added with their low surface brightness (particularly for high-redshift galaxies) this diversity can cause various systematic biases in their detection and photometry. We introduce a new noise-based method to detect and segment galaxies deeply drowned in noise [1]. It imposes statistically negligible constraints on the to-be-detected targets. We are able apply a sub-sky threshold (roughly equivalent to -0.5 sigma) to the image for the first time. This allows for very accurate non-parametric detection of the low surface brightness structure in the outer wings of bright galaxies or the intrinsically faint objects that remain wholly below the commonly used thresholds (>1 sigma). Both these targets play a crucial role in our understanding of galaxy formation.

The false detections are identified and removed using the ambient noise as a reference, thereby achieving a purity (fraction of true to the total number of detections) of 0.88 as compared to 0.29 for SExtractor when completeness (fraction of true to total number of mock profiles) is 1 for a sample of extremely faint and diffuse identical mock galaxy profiles. The dispersion in their measured magnitudes is less by one magnitude. By defining the accuracy of detection as the difference of the measured sky with a known background of mock images, 4.6 times less biased sky measurement is achieved depending on the diffuseness of the sources. Contrary to the existing signal-based approach to detection, in its various implementations, signal-related parameters such as the image

point spread function or known object shapes and models are irrelevant here. NoiseChisel is our software implementation of this noise-based algorithm, it is distributed as part of the GNU Astronomy Utilities (Gnuastro) [2]. Gnuastro is the first astronomical software that fully conforms with the GNU Coding Standards and thus integrates nicely with Unix-like operating systems, while having a standardized coding style, familiar command-line user interface, and a comprehensive manual in various web-based, print and command-line formats [3]. All the data-generated numbers and plots in this work are exactly reproducible with a publicly released reproduction pipeline containing all the scripts and configuration files managed through a Makefile [4].
*[1] Akhlaghi, M. and T. Ichikawa (2015). ApJS, 220, 1 (arXiv:1505.01664), [2] https://www.gnu.org/software/gnuastro/, [3] https://www.gnu.org/software/gnuastro/manual/ [4] https://gitlab.com/makhlaghi/NoiseChisel-paper/*

---

**O2.4 Tony Butler-Yeoman**
Victoria University of Wellington

*Detecting Diffuse Sources in Astronomical Images*
*joint work with Marcus Frean, David W. Hogg, and Chris Hollitt*

We present an algorithm capable of detecting diffuse, dim sources of any size in an astronomical image. These sources often defeat traditional methods, which expand regions around points of high intensity. Extended sources often have no bright points and are only detectable when viewed as a whole, so a more sophisticated approach is required. Our algorithm operates at all scales simultaneously by considering a tree of nested candidate bounding boxes, and inverts a hierarchical Bayesian generative model to obtain the probability of sources existing at given locations and sizes. This model naturally accommodates the detection of nested sources, and no prior knowledge of the distribution of a source, or even the background, is required. The algorithm scales linearly with the number of pixels, is feasible to run on a full-sized SKA image, and requires minimal parameter tweaking to be effective. We demonstrate the algorithm on several types of astronomical and artificial images, and show that (on tested data) the algorithm can detect most significant sources, with a low rate of false positives. This paper is accompanied by an open-source reference implementation in Java.

---

**O2.5 Tianheng Liang**
College of Information Science and Technology

*A Modified Method of Extracting Filaments from Astronomical Images*

Filaments are a type of wide-existing astronomical structure. It is a challenge to distinguish filamentary structures from background images because filamentary radiation intensity is usually weak and filaments often mix with bright objects, e.g. stars, which leads difficulty to separate them. In 2013, A. Menâ shchikov proposed a multi-scale, multi-wavelength filament extraction method which was used to extract filaments from bright sources, noise, and isotropic background. The method decomposes a simulation astronomical image containing filaments into spatial scale images

to prevent interaction influence of different spatial scale structure. However, the algorithm of processing each single spatial scale image in the method is employed to simply remove tiny structures by counting connected pixels number. Removing tiny structures based on local information only might remove some part of the filaments because filaments in real astronomic image are usually weak. We tempt to use MCA (Morphology Components Analysis) in order to process each singe spatial scale image. MCA uses a dictionary whose elements can be wavelet translation function, curvelet translation function or ridgelet translation function to decompose images. Different selection of elements in dictionary can get different morphology components of the spatial scale image. By using MCA, we can get line structure, gauss sources and other kind of structures in spatial scale images and exclude the components that are not related with filaments. Our experiments show that our method is efficient in filaments extraction from real astronomic images.

---

## SESSION III: LSST AND LESSONS LEARNED FROM CURRENT PROGRAMS

### O3.1 Mario Juric

University of Washington

*LSST: Building the Data System for the Era of Petascale Optical Astronomy*

The Large Synoptic Survey Telescope (LSST; http://lsst.org) is a planned, large-aperture, wide-field, ground-based telescope that will survey half the sky every few nights in six optical bands from 320 to 1050 nm. It will explore a wide range of astrophysical questions, ranging from discovering “killer” asteroids, to examining the nature of dark energy.

The LSST will produce on average 15 terabytes of data per night, yielding an (uncompressed) data set of over 100 petabytes at the end of its 10-year mission. To enable the wide variety of planned science, the LSST Project is leading the construction of a new, general-purpose, high-performance, scalable, well documented, open source data processing software stack for O/IR surveys. Prototypes of this stack are already capable of processing data from existing cameras (e.g., SDSS, DECam, MegaCam), and form the basis of the Hyper Supreme-Cam (HSC) Survey data reduction pipeline. In the 2020-ies, running on dedicated HPC facilities, this system will enable us to process the LSST data stream in near real time, with full-dataset reprocessings on annual scale.

In this talk, I will review the science goals and the technical design of the LSST, focusing on the data management system, its architecture, the software stack, and the products it will generate. I will discuss the exciting opportunities it presents for LSST, and the astronomical software community as a whole. More broadly, I will also discuss implications of petascale data sets for astronomy in the 2020s, and ways in which the astronomical community can prepare to make the best use of them.

---

**O3.2 Christian Wolf**
RSAA Mt Stromlo ANU

*Developing Data Processing Pipelines for Massive Sky Surveys - Lessons Learned from SkyMapper*

The SkyMapper Survey led by Mt Stromlo Observatory at ANU started in 2014 to map the Southern sky in six spectral passbands. It will identify over 100 million celestial objects and be the first deep and detailed digital resource covering the entire Southern sky. For several years, its Science Data Pipeline (SDP) has been developed. Here we report on lessons learned from this ongoing effort. Should you be ready when the data flow in? How reliably can the required functionality be anticipated? What is the role of staff turnover in a long-term development? What issues are posed by dependencies? And, is a supercomputer the best platform?

---

**O3.3 Jesus Salgado**
ESAC/ESA

*Access to Massive Catalogues in the Gaia Archive: a New Paradigm*

New astronomical missions have reinforced the change on the development of archives. Archives, as simple applications to access the data are being evolving into complex data center structures where computing power services are available for users and data mining tools are integrated into the server side.

In the case of astronomy science that involves the use of big catalogues, as in Gaia or Euclid, the common ways to work on the data need to be changed to a new paradigm "move code close to the data", what implies that data mining functionalities are becoming a must to allow the science exploitation.

Some massive operations like cross match between catalogues, integration of big queries into workflows by serialization of intermediate results in cloud resources like VOSpace, integration of data mining tools in virtualized environments, etc are being integrated into the ESAC Gaia archive. Also, totally new science use cases like, e.g., asteroids discovered by Gaia, combine astronomy with small bodies science and the results should be made available in a transparent way to different communities, allowing filtering, clustering and, in general, integrated data mining techniques.

We present the tools already available in the Gaia archive for big catalogues manipulation, like cross match operations, pre-cooked cross match tables with the main astronomical catalogues, and, also, the ongoing work on the publication of a huge variety of data objects, e.g. asteroids, and other add-ons that would allow scientists a new way to produce science.

---

## KEYNOTE ADDRESS

**Brian Schmidt**
Australian National University

*Big Data and Big Astronomy*

Astronomy is considered a pioneering field in the area of e-research and Big Data, with the 25 ADASS conferences testament to this activity. Several of the most important discoveries of the past two decades have been enabled by the sophisticated analysis of very large datasets (in the context of other disciplines at the same time). Many of the upcoming telescopes and their flagship science programs have built into them massive data processing elements. I will provide an overview of the scientific motivation for why the work in Astronomical Data Analysis Software & Systems is so important.

---

## SESSION IV: KNOWLEDGE DISCOVERY AND DATA MANAGEMENT TOOLS FOR ASTRONOMICAL BIG DATA

### O4.1 Steven Berukoff

National Solar Observatory

*Petascale Data Management in Solar Physics: Approach of the DKIST Data Center*

When construction is complete in 2019, the Daniel K. Inouye Solar Telescope will be the most-capable large aperture, high-resolution, multi-instrument solar physics facility in the world. The telescope is designed as a four-meter off-axis Gregorian, with a rotating Coude laboratory designed to simultaneously house and support five first-light imaging and spectropolarimetric instruments. At current design, the facility and its instruments will generate data volumes of 5 PB, produce 10^8 images, and 10^9 metadata elements annually. This data will not only forge new understanding of solar phenomena at high resolution, but enhance participation in solar physics and further grow a small but vibrant international community.

The DKIST Data Center is being designed to store, curate, and process this flood of information, while augmenting its value by providing association of science data and metadata to its acquisition and processing provenance. In early Operations, the Data Center will produce, by autonomous, semi-automatic, and manual means, quality-controlled and -assured calibrated data sets, closely linked to facility and instrument performance during the Operations lifecycle. These data sets will be made available to the community openly and freely, and software and algorithms made available through community repositories like Github for further collaboration and improvement.

We discuss the current design and approach of the DKIST Data Center, describing the development cycle, early technology analysis and prototyping, and the roadmap ahead. In this budget-conscious era, a key design criterion is elasticity, the ability of the built system to adapt to changing work volumes, types, and the shifting scientific landscape, without undue cost or operational impact.

We discuss our iterative development approach, the underappreciated challenges of calibrating ground-based solar data, the crucial integration of the Data Center within the larger Operations lifecycle, and how software and hardware support, intelligently deployed, will enable high-caliber solar physics research and community growth for the DKIST's 40-year lifespan.

---

**O4.2 Kyler Kuehn**
Australian Astronomical Observatory

*Managing the Data Deluge from the Dark Energy Survey*

The Dark Energy Survey comprises observations over 5000 square degrees of the southern hemisphere. Its key science goal is the measurement of the time-dependent and time-independent components of the dark energy equation of state using four probes: weak gravitational lensing, galaxy clusters, large-scale structure, and Type Ia supernovae. The 570 Megapixel Dark Energy Camera used for this survey produces several GB of raw data with every image (approximately every 2 minutes during observations). The entire DES dataset collected over five years of Survey operation will be of order 1 Petabyte in size. Significant resources have been devoted to the management and analysis of DES data, including the development of numerous hardware and software applications for near-real-time processing. We describe the general process of DES Data Management, along with selected examples of some of the scientifically productive "end-user" tools.

---

**O4.3 Lloyd Harischandra**
Australian Astronomical Observatory

*Hadoop and Spark for Data Management, Processing and Analysis of Astronomical Big Data: Applicability and Performance*

The AAT node for the All Sky Virtual Observatory (ASVO) is being built on top of Apache Hadoop and Apache Spark technologies. The Hadoop Distributed File System (HDFS) is used as the data store and Apache Spark is used as the data processing engine. The data store consists of a cluster of 4 nodes of which 3 nodes provide space for data storage and all 4 nodes can be used to gain computing power. In this talk, we compare the performance of Apache Spark on GAMA data hosted on HDFS against other relational database management systems and software in the fields of data management, real-time processing and analysis of astronomical Big Data.

We examine the usability, flexibility and extensibility of the libraries and languages available within Spark, specifically in querying and processing large amounts of heterogeneous astronomical data. The data included are primarily in tabular format but we discuss how we can leverage the rich functionalities offered by Hadoop and Spark libraries to store, process/transform and query data in other formats such as HDF5 and FITS. We will also discuss the limitations of existing relational database management systems in terms of scalability and usability.

Then we evaluate the benchmark results of varying data import and transform scenarios, and the expected latency of queries across a range of complexities.

Lastly, we will show how astronomers can create custom data-processing tasks in their preferred language (python, R etc.) using Spark, with limited knowledge of the Hadoop technologies

---

**O4.4 Alberto Accomazzi**
Harvard-Smithsonian Center for Astrophysics

*Aggregation and Linking of Observational Metadata in the ADS*

We discuss current efforts behind the curation of observing proposals, archive bibliographies, and data links in the NASA Astrophysics Data System (ADS). The primary data in the ADS is the bibliographic content from scholarly articles in Astronomy and Physics, which ADS aggregates from publishers, arXiv and conference proceeding sites. This core bibliographic information is then further enriched by ADS via the generation of citations and usage data, and through the aggregation of external resources from astronomy data archives and libraries. Important sources of such additional information are the metadata describing observing proposals and high level data products, which, once ingested in ADS, become easily discoverable and citeable by the science community. Additionally, bibliographic studies have shown that the integration of links between data archives and the ADS provides greater visibility to data products and increased citations to the literature associated with them. While ADS solicits and welcomes the inclusion of this observational metadata from all astronomy data centers, the curation of bibliographies, observing proposals and links to data products is left to the archives which host the data and which have the expertise and resources to properly maintain them. In this respect, the role of the ADS is to provide tools and services to facilitate the creation and maintenance of these resources. We describe these curation workflows, recommending the adoption of best practices in support of such data curation and management activities and discuss how these efforts relate to broader cross-disciplinary data citation initiatives.

---

## SESSION V: DATA PIPELINES

**O5.1 Janet Evans**
Smithsonian Astrophysical Observatory

*The Chandra Souce Catalog Pipeline: the Yin Yang of a Challenging Project*

Production of Release 2 of the Chandra Source Catalog (CSC2) started in Spring of 2015. Development of the processing pipeline is the work of a small team of Chandra scientists and software developers each contributing their knowledge and expertise to the requirements, design, and implementation of the pipeline, tools, and archive needed to produce the catalog. CSC2 will

triple the size of the original catalog released in 2010 to ~300,000 sources. The source increase is due to co-adding multiple observations and the use of new source detection and background algorithms to include the faintest (~5 net counts) sources. The images and source detections of the deepest and most crowded fields are spectacular, and very faint sources will be found! CSC2 is innovative in that data products (e.g., images, exposure maps) are saved in the archive, along with tabular data (e.g., source positions, position error) for user access and analysis. The pipeline goes well beyond the traditional data reduction thread and well into detailed data analysis. We had to find the balance in developing a pipeline that runs 24/7 against the intricacies of analysis required for source detection of co-added fields. We had to evaluate where to draw the scientific results line since it impacted the project in both time and resources. We had to measure hardware costs and estimated computation time against affordability and project pressures to complete the catalog. In this paper, we will provide an overview of the technical challenges we met in developing the CSC2 pipeline, and review the end products available to the catalog user. We will highlight the coordination and management challenges in developing the catalog against responsibilities of a team in the out years of a mission. We will highlight lessons learned - both positive and negative so they can benefit others on a similar path.

---

**O5.2 Martin Kuemmel**
Ludwig-Maximilians-Universitaet Muenchen

*Data Challenges for the Euclid Cataloging Pipeline*

*Co-Author: S. Pilo, M. Castellano, A. Boucaud, A. Fontana, H. Dole, P. Atreya, R. Cabanac, J. Coupon, S. Desai, P. Guillard, R. Henderson, J. Mohr, S. Paltani, P. Petrinca, M. Wetzstein*

Euclid is an ESA mission to be launched in 2020. The satellite will determine the expansion rate of the Universe at various cosmic ages with an unpreceded accuracy by measuring Weak Gravitational Lensing and Galaxy Clustering up to z~2. Euclid will observe 15,000 deg^2 with its two instruments, the Visible Imaging Channel (VIS) and the Near IR Spectrometer and imaging Photometer (NISP). The satellite data is completed with ground based griz imaging from surveys such as the Dark Energy Survey to enable photo-z determination.

Generating the the multiwavelength catalog of Euclid and ground based data is a central part of the entire Euclid data reduction pipeline. On behalf of OU-MER, the unit of the EUCLID Science Ground Segment responsible for this task, I will discuss the concepts and strategies to generate the Euclid catalogues that meet the tight requirements on photometric accuracy. While the object detection had been presented in last year's ADASS, the main focus here are the procedures to estimate the photometry on images with different depths and resolutions (0.2 "for VIS data and ~1.0" for ground based data), which were developed for deep surveys such as CANDELS or GOODS.

We will present the results of our cataloging procedures on emulated Euclid data assembled in two Data Challenges in the GOODS-South field (0.05 Deg^2, from HST imaging) and the COSMOS field (~1deg^2, from HST and ground based imaging). The results from photo-z, obtained by the unit OU-PHZ on the emulated data, are shown and the first concepts for the verification and validation of our products are discussed.

**O5.3 Carlos Gabriel**
ESA/ESAC

*The XMM-Newton Pipeline Processing System PPS and the 3XMM-DR5, the Largest-Ever Catalogue of X-ray Sources*

The European Space Agency X-ray space observatory XMM-Newton has executed ~600 observations per year since January 2000. Science data from the 3 imaging and 2 grating X-ray instruments and the UV/optical telescope on XMM-Newton (which observe simultaneously) are reduced by a dedicated pipeline processing system (PPS), using the same Scientific Analysis System (SAS) software packages that are available for users to interactively analyse XMM-Newton data. The pipeline, originally developed and maintained during the first 12 years of the mission by the Survey Science Centre (SSC), a Consortium of European institutes, is now operated by, and under the direct responsibility of, the XMM-Newton Science Operations Centre at ESAC.

Among the many (~500) products derived for each observation (from calibrated event lists to individual and combined sky images, diagnostic images, cross correlations with archival data, spectra, time series, etc) are the lists of detected sources in the field of view (typically 50-100 per observation). From these source lists several source catalogs have been compiled during the mission, each of them marking a new record in number of objects. The fifth release of the XMM-Newton serendipitous source catalogue (3XMM-DR5), made public by the SSC in April 2015, contains 565.962 X-ray detections, with 396.910 unique sources, ranging from nearby objects in our Solar System to supermassive black holes at the edge of the Universe. For each detection, a wealth of information is provided to help understand the nature of the object, including more than 133.000 spectra and time series of individual detections.

We are going to discuss in this contribution the continuous development and maintenance of the pipeline, including its move from the University of Leicester to ESAC, as well as the characteristics and potential of the catalogue, and the technical challenges for building it.

---

**O5.4 Laurence Chaoul**
CNES

*Processing Gaia's Billion Stars, a Big Data Story*

ESA's Gaia astronomy mission, that was launched 19th December 2013, aims at mapping more than one billion stars and objects. The scientific data processing has been delegated to the Data Processing and Analysis Consortium (DPAC), which is composed of more than 400 people all across Europe. This consortium, organized in 9 scientific coordination units across 6 data processing centers, is presented herein. Responsible for a data processing centre (DPCC) that will run 3 scientific coordination units, CNES is in charge of a large part of the Gaia data processing. DPCC has to answer in two main technical challenges: a huge data volume to handle (PetaByte order) and complex processing algorithms with timeliness constraints. Since 2006, scientists have developed Java

algorithms to be integrated on a prototype architecture based on PostgreSQL and a Torque/Maui distributed resources manager / scheduling system. In 2010, the initial test campaigns proved the inability of this design to meet the performance requirements. A study aimed at finding an alternative was then launched with the strong constraint not to break the existing interface with the scientific algorithms. Finally, CNES with Thales has chosen a solution based on Hadoop technology, emerging from the internet applications such as Facebook or Ebay, to supersede not only PostgreSQL but also Torque/Maui. Hadoop efficiency is due to several reasons: its distributed File System (HDFS) with file locality, its distribution mechanism where the processing goes to the data, and finally MapReduce (MR), a very powerful data manipulation paradigm. Another higher level efficient framework was selected to simplify the data manipulation queries writing: Cascading. Moreover Thales uses PHOEBUS CNES product to manage the high level orchestration need between different scientific algorithms. All software components are deployed on a cluster composed of high density nodes. Each node provides both computing power and storage space. The DPCC design is entirely scalable: cluster performances are linearly linked to the nodes number. The cluster is currently composed of 1152 processor cores, and it will grow according to the need. Thanks to this scalability, CNES can benefit from the last technological breakthroughs (cores and hard disks) in every new hardware supply, and easily adapt its computing power. Current estimations give 6000 processor cores (meaning about 500 servers) required to process all data at the end of the mission After few months of daily processing, a complete feedback on cluster performances has been made to conclude that the CPU is the current bottleneck. As a consequence, a more powerful processor has been chosen for the mid-2015 supply. The current Hadoop software version has also been updated, in order to take advantage of YaRN, also known as MRv2 (more scalable, more efficient and more flexible). After a brief description of the Gaia project and of the CNES involvement in the Gaia data processing, the discussion will present the DPCC design around Hadoop technology, by focusing mainly on the first results and performances on the executions of the spectroscopic scientific chain, run every day in the DPCC. The web portal “GaiaWeb based on the ElasticSearch technology-, allowing scientists to follow operation progress and to value the huge volume of DPCC produced data for further analysis will be also succinctly mentioned. We will conclude with the perspectives and lessons learned on scientific big data processing for Gaia and other space projects.

---

## SESSION VI: REAL TIME PROCESSING

### O6.1 Mike Wise
**ASTRON**

*Radio Astronomy Data Access Levels Up: From the LOFAR Archive to SKA Data Centres*

The Big Data challenge has become a pervasive theme in Astronomy in recent years. In an era of new, large-scale astronomical facilities such as the SKA, CTA, and LSST, the typical sizes of the datasets that researchers must interrogate has already reached the petabyte-scale with data collections well on their way to breaking the exabyte barrier. While certainly not unique to radio astronomy, new technology radio facilities like LOFAR, the MWA, LWA, ASKAP, MeerKAT are already

pushing the boundaries for our current data management and analysis technologies. For example, with a current data archive of over 20 petabytes, LOFAR already has one of the largest radio astronomy data collections in the world. Data at these scales present unique challenges not just for managing the collection but also for how researchers extract their science. More positively, precursor instruments like LOFAR, provide an opportunity to tackle these problems today in order to prepare for next generation of instruments like the SKA.

In this talk, I will describe the current LOFAR data flow system and long-term archive, its current capabilities, and how it functions as an integrated part of the operational telescope. I will also discuss ongoing efforts to expand that capability beyond basic pipeline processing to include more analysis and science extraction functionality. This evolution from traditional store-and-retrieve archive to flexible analysis environment is a test bed for the development of Regional Science Data Centres for the SKA. A global network of such regional SDCs are currently envisioned to enable the community to take maximal advantage of the scientific potential of the SKA and will likely be the main working interface for most scientists using SKA data. I will conclude by describing current efforts within the SKA community to define and establish a network of such regional SDCs.

---

**O6.2 Jan David Mol**
ASTRON

*COBALT: Replacing LOFAR's on-line Correlator & Beamformer*

Replacing a major component in a production system using new technology carries inherent risks. Yet this was the goal of the COBALT project: to replace LOFAR's on-line correlator & beamformer, that ran on a supercomputer, with a GPU-based cluster built using off-the-shelf hardware.

This fundamental change in platform required a nearly complete rewrite of the correlator software, on top of parallel platforms such as CUDA, OpenMP, and MPI. COBALT needed to be a drop-in replacement, with limited opportunity for running alongside to the old system. It therefor faced hard requirements with respect to its performance, correctness, scalability, and its delivery deadline.

COBALT successfully runs as LOFAR's on-line processor. Our development and commissioning methods allowed us to develop features that were correct, robust, and performant, without introducing regression. Project planning allowed us to choose at early stages which features could and should be delivered on time.

We will present (1) an overview of the COBALT system (2) the development methods that we used that were key to meet our requirements, and (3) the pitfalls and surprises we faced.

---

**O6.3 Nuria Lorente**
Australian Astronomical Observatory

*Path-Finding Algorithms for TAIPAN's Starbug Robots*

The AAO's TAIPAN instrument deploys 150 8mm diameter Starbug robots to position optical fibres to accuracies of 0.5 arcsec, on a 30 cm glass field plate on the focal plane of the 1.2 m UK-Schmidt telescope. This paper describes the software system developed to control and monitor the Starbugs, with particular emphasis on the automated path-finding algorithms, and the metrology software which keeps track of the position and motion of individual Starbugs as they independently move in a crowded field. The software employs a tiered approach to find a collision-free path for every Starbug, from its current position to its target location. This consists of three path-finding stages of increasing complexity and computational cost. For each Starbug a path is attempted using a simple method. If unsuccessful, subsequently more complex (and expensive) methods are tried until a valid path is found or the target is flagged as unreachable. The challenge is to ensure that a given path takes a minimum amount of time to execute, but that doing this does not prevent a neighbouring Starbug from reaching its target. Additionally, the algorithm must ensure that the umbilical which tethers each Starbug to the instrument (and carries the science fibre) is not entangled with that of other Starbugs in their simultaneous movement across the field plate, as this would increase the risk of reconfiguration failure in subsequent exposures. Simulations show that this multi-stage approach allows the field to be reconfigured within the required 5 minutes for the majority of expected target configurations. We will show the results of initial tests with the instrument using the recently-completed production Starbugs.

---

**O6.4 Matthew Bailes**
Swinburne University of Technology

*Real-Time Searching for Fast Radio Bursts and other Radio Transients Using the UTMOST Telescope*

A new class of radio source known as fast radio bursts (FRBs) was identified at the Parkes radio telescope and presented by Lorimer et al. (2007). Since then two other observatories have detected FRBs which are characterised by very short durations (a few milliseconds) and exhibit the dispersion sweep characteristic of celestial sources.

The FRBs occur relatively infrequently with less than 20 currently known in over 8 years of searches. To accelerate the rate of discovery we have designed and implemented a novel software correlator that takes 22 GB/s of radio data and forms 700 fan beams to search 8 square degrees for FRBs in real time using 250 Teraflops of GPUs at the 18,000 m^2 Molonglo telescope. The project is dubbed the "UTMOST". To approach system design sensitivity novel radio frequency interference algorithms have been implemented as the telescope operates in the mobile phone band. Whilst searching for FRBs the telescope can also time pulsars, write filterbank data to disk for pulsar surveys and make interferometric maps. The interferometric nature of the array allows it to determine the "parallax" of any source to determine whether it is terrestrial or not so the instrument is very adept at

separating genuine celestial signals from interference. The computational challenges and early results will be presented.

---

**O6.5 Ewan Barr**
Swinburne University of Technology

*Massive data streaming and processing for pulsar timing*

The improved sensitivity, flexibility and survey speed of the next generation of radio telescope arrays (MeerKAT, ASKAP, SKA) comes at the cost of a vast increase in the amount of data that must be captured and analysed. This is particularly true for the Square Kilometre Array (SKA), where in Phase 1 the mid-frequency antennas will produce up to 28 Tb/s of data. Such data rates make storage and offline processing unfeasibly expensive, thus necessitating the development of real-time processing hardware and software that can reliably identify and preserve data of scientific interest.

For the SKA, the Central Signal Processor (CSP) will handle real-time processing of observations. This large hybrid FPGA-GPU supercomputer will perform intensive tasks such as RFI rejection, polarisation calibration, correlation, beamforming, pulsar searching and pulsar timing. At Swinburne University of Technology we are working on the design and development a pulsar timing instrument that will facilitate one of the SKA's key science goals: "Strong-field tests of gravity using pulsars and black holes". This instrument will capture and process 1.2 Tb/s of beamformed data, performing interference removal, mitigating of interstellar medium propagation effects, channelization and phase folding. To do this we use off-the-shelf graphics cards and high-speed network interfaces to produce a new pulsar timing instrument an order of magnitude more powerful than the current generation.

In this talk I will discuss the role of pulsar timing in the SKA and review the key design aspects of our instrument, highlighting the computational challenges involved and the constraints due to budget, power and environment.

---

## SESSION VII: KNOWLEDGE DISCOVERY AND DATA MANAGEMENT TOOLS FOR ASTRONOMICAL BIG DATA

**O7.1 Ann Marie Cody**
NASA Ames Research Center

*Multiwavelength Variability Surveys: Reaping the Stellar Harvest*

Over the past five years, a number of dedicated stellar variability surveys have launched from both the ground and space. Many of these programs focus on the detection of specific events, such as exoplanet transits or extragalactic transients. Yet the observed variability behavior encompasses a

much larger range of stellar phenomena. To take full advantage of variability survey data, we must detect and classify distinct morphological features in light curves. This task has been particularly challenging for the young (1-10 million year old) stars, which are well known to vary at the 1-100% level on timescales of hours to years. In this talk, I will highlight recent progress in the identification, classification, and physical understanding of young star variability. I will present a selection of optical and infrared time series of pre-main sequence stars and brown dwarfs from state-of-the-art datasets, including the Young Stellar Object Variability (YSOVAR) Campaign with the Spitzer Space Telescope. I will describe the data storage approaches and time series analysis techniques employed to extract physically meaningful information from the light curves. The lessons learned from YSOVAR and other campaigns should be broadly applicable to massive future surveys such as TESS and LSST.

---

**O7.2 Amr Hassan**
Swinburne University of Technology

*Enabling Science in the Petascale Era: MWA Data Archive and Dissemination Platform*

Current and planned large-scale astronomical facilities will boost our ability to gather raw data from the universe but this does not necessarily mean they will increase our knowledge at the same rate. The ability to transform massive raw data into usable information is a challenge that astronomers will have to face in their day-to-day activities. Providing the majority of astronomers with the ability to access, analyze, visualize, and process this raw data and its derived data products is a vital step towards better utilization of this data.

The Murchison Widefield Array (MWA) is the low frequency precursor for the Square Kilometre Array (SKA). It has been operational since July 2013 producing around 6 TB/day of raw visibility data. It is one of the few operational Petascale astronomical facilities. Using MWA as the main case study, this work discusses the challenges that face radio astronomy in the Petascale data era. We will illustrate why these challenges cannot be approached in a “business-as-usual” evolutionary manner. As a pointer to the way forward, we will then introduce a new data archive and dissemination platform that aims to provide national and international researchers with seamless access to different MWA data products via online services and tools.

---

**O7.3 Paul Hirst**
Gemini Observatory

*A New Data Archive for Gemini - Fast, Cheap, and in the Cloud*

We have deployed a new data archive for Gemini Observatory, taking less than 2 years and 3 FTE from project start to public deployment, and under strong budget constraints. In doing so, we have used several novel techniques which allowed rapid development while providing versatile search and download functionality alongside more advanced features such as calibration association, low latency on new data availability and various forms of API access to support the needs of our user community.

The new archive system shares the same software code base as our in house data management tools, and is deployed on Amazon Web Services.
In this presentation I will describe the more novel aspects of the architecture and some of the powerful features that these enable, and the former enabled the latter with a very modest development effort.

---

**O7.4 Fabio Pasian**

INAF - Osservatorio Astronomico di Trieste

*ASTERICS - Addressing Cross-Cutting Synergies and Common Challenges for the Next Decade Astronomy Facilities*

*Authors: Fabio Pasian, Michael Garrett, Francoise Genova, Giovanni Lamanna, Stephen Serjeant, Arpad Szomoru, Rob van der Meer*

The large infrastructure projects for the next decade will allow a new quantum leap in terms of new possible science. ESFRI, the European Strategy Forum on Research Infrastructures, a strategic instrument to develop the scientific integration of Europe, has identified four facilities (SKA, CTA, KM3Net and E-ELT) to support. ASTERICS (Astronomy ESFRI & Research Infrastructure Cluster) aims to address the cross-cutting synergies and common challenges shared by the various Astronomy ESFRI facilities and other world-class facilities. The project (22 partners across Europe) was funded by the EU Horizon 2020 programme with 15 MEuro in 4 years. It brings together for the first time, the astronomy, astrophysics and particle astrophysics communities, in addition to other related research infrastructures.

The major objectives of ASTERICS are to support and accelerate the implementation of the ESFRI telescopes, to enhance their performance beyond the current state-of-the-art, and to see them interoperate as an integrated, multi-wavelength and multi-messenger facility. An important focal point is the management, processing and scientific exploitation of the huge datasets the ESFRI facilities will generate. ASTERICS will seek solutions to these problems outside of the traditional channels by directly engaging and collaborating with industry and specialised SMEs. The various ESFRI pathfinders and precursors will present the perfect proving ground for new methodologies and prototype systems. In addition, ASTERICS will enable astronomers from across the member states to have broad access to the reduced data products of the ESFRI telescopes via a seamless interface to the Virtual Observatory framework. This will massively increase the scientific impact of the telescopes, and greatly encourage use (and re-use) of the data in new and novel ways, typically not foreseen in the original proposals. By demonstrating cross-facility synchronicity, and by harmonising various policy aspects, ASTERICS will realise a distributed and interoperable approach that ushers in a new multi-messenger era for astronomy. Through an active dissemination programme, including direct engagement with all relevant stakeholders, and via the development of citizen scientist mass participation experiments, ASTERICS has the ambition to be a flagship for the scientific, industrial and societal impact ESFRI projects can deliver.

---

## SESSION VIII: ALGORITHMS FOR ASTRONOMICAL DATA REDUCTION

**O8.1 Thomas Robitaille**
Max Planck Institute for Astronomy

*The Astropy Project: Current Status and Future Plans*

The Astropy Project is a community effort to develop a single core package for Astronomy in Python and foster interoperability between Python Astronomy packages, and is one of the largest open-source collaborations in Astronomy. In this talk I will present an overview of the project, provide an update on the latest status of the core package, which saw the v1.0 release this year, and outline our plans for the coming year. In addition, I will describe the "affiliated packages": Python packages that use Astropy and are associated with the project, but are not actually a part of the core library itself, and will give an overview of the tools we have made available to allow anyone to develop their own domain-specific affiliated package.

---

**O8.2 Greg Madsen**
Cambridge Astronomy Survey Unit (CASU), University of Cambridge

*An Infrared Search for Satellite Orbital Debris*

Less than 20% of the more than 15,000 objects currently in orbit around the Earth are operational. The number, size distribution, and location of orbital space debris are very poorly known. Collisions with space debris are a major risk to the increasing number of operational satellites that modern society relies upon. We describe new efforts to find and characterise orbital debris with the infrared camera (WFCAM) on the 4-meter class UKIRT telescope. We discuss algorithms for identifying orbiting objects in sidereal and non-sidereal tracked images and for calibrating their astrometry and photometry. We highlight our results to date which include a large survey of the geosynchronous belt, the dependence of IR colours on solar phase angle, and high time resolution light curves.

---

**O8.3 Elise Hampton**
Research School of Astronomy & Astrophysics, Australian National University

*Using an Artificial Neural Network to Classify Multi-Component Emission Line Fits*

Integral Field Spectroscopy (IFS) is changing our approach to the study of galaxy evolution. Surveys such as CALIFA (Sanchez et al. 2012), SAMI (Croom et al. 2012), MANGA (Bundy et al. 2015), and S7 (Dopita et al. 2014) are building databases of hundreds to thousands of galaxies ready to explore galaxy evolution as a function of morphological and spectroscopic classification. Access to this

information comes at a price: data volume. Data reduction pipelines address the problems of preparing data for analysis but understanding the contents of these data cubes remains a significant challenge. Automated continuum and absorption line fitting is routinely used to understand the stellar populations in galaxies, while emission line fitting provides insight into active star formation, AGN activity and shock properties of galaxies. This type of pre-analysis is time consuming for IFU surveys and is no longer feasible by hand as we understand that there can be multiple processes behind an emission line. Automated emission line fitting, including multiple components (e.g. Ho et al. in press), are currently in use but still requires human input to decide the best number of components to describe each emission line. This presentation describes our automated machine learning algorithm to remove this time consuming human input and streamline multi-component emission line fitting for surveys. We have taken what was years of work by a person to hours of computer time using an artificial neural network.

---

**O8.4 Ian Stewart**
Sterrewacht Leiden

*LIME - the Line Modeling Engine*

The Atacama Large Millimetre/submillimetre Array (ALMA) has begun to produce spectacular images showing the distributions of molecular excited states in a range of different cosmic environments. Understanding what is going on in these environments can be difficult however, partly because the objects are often optically thick, and partly because they may have complicated three-dimensional structures which are difficult to interpret given the sparse clues available from the observation. A helpful technique is forward modelling, in which the equations controlling the balance between radiation density and populations of excited states are solved for a chosen model, from which the expected appearance of the object in an ALMA observation can be predicted. LIME is a package for solving these equations on an adaptively-weighted grid of points for a three-dimensional object of arbitrary configuration. LIME can be used stand-alone but is rendered much more powerful if run inside the wrapper package ARTIST, which provides many features including a GUI, a convenient viewer, and access to a library of model templates. In this talk I describe recent advances in both packages, which include improvements to the solving and raytracing algorithms, integration of ARTIST within CASA, and a new interface to LIME which allows it to be more easily run from within a pipeline. The talk will be illustrated with examples of successful applications of LIME to ALMA images.

---

## SESSION IX: DATA PIPELINES

**O9.1 Sarah Hegarty**
Swinburne University of Technology

*Realistic Imaging of Simulated Universes with the Theoretical Astrophysical Observatory*

The astronomical Big Data era will deliver a diversity of rich data sets: while observational astronomers plan the next generation of all-sky surveys, theorists are developing ever larger and more sophisticated cosmological simulations. Increasingly, though, working with these specialised data products demands specialised expertise - making it increasingly challenging for the astronomer to compare observation and theoretical prediction. We describe a data pipeline designed to facilitate such comparisons, bridging the gap between observers and theorists in the astronomical community.

Our pipeline extends the functionality of the Theoretical Astrophysical Observatory (TAO). TAO is an online virtual laboratory which couples galaxy formation models to a suite of large cosmological simulations, to deliver simulated galaxy survey catalogues. To date, these simulated survey catalogues have not been accompanied by imaging data products, limiting their usefulness to the observer.

A number of software packages have been developed to produce such simulated imaging data; however, most are standalone, and lack integration with a dedicated object catalogue generator. Accordingly, we have investigated, adapted, and integrated several image simulation packages for use with TAO, and have developed a new “virtual telescope” pipeline which delivers highly realistic imaging of simulated galaxy surveys. Easily accessible from a clean and intuitive web interface, this new capability opens up new horizons for cross-disciplinary astronomical knowledge discovery in the cloud.

---

**O9.3 Hadrien Devillepoix**
Desert Fireball Network, Curtin University

*Handling Ultra-Wide Field Astronomical Data: The Story of a Continent-Scale Observatory*

Years of development in consumer camera technology have made digital cameras sensors a viable option for sky observation, at a much cheaper price than dedicated CCD technology. The Desert Fireball Network (DFN) is building and operating a continent scale distributed robotic observatory using this technology (34 units currently deployed, covering 1.5kmÂ²). The cameras consist of 36 Mpixels DSLRs with fish-eye lenses, fireball detections are corroborated with other cameras, and the resulting triangulation yields meteorite fall site and orbit in the solar system. The full data is archived on hard drives for later retrieval and further analysis. While using fish-eye lenses is convenient for seeing the whole sky on a single camera, doing astrometry on those lenses is a big issue. And precise

astrometry is the only way to get a reasonable search area for meteorites. Available software cannot cope with the strong distortion. We present here a new method for calibrating fish-eye lenses, based on a computational iterative polynomial fitting process. A more precise result can be obtained on a cropped region of the image, combining multiple images. This method is particularly useful to get maximum precision on a particular event of interest. While the DFN is designed to triangulate near-field bright objects, a small upgrade to this system (50mm lens instead of fish-eye) can make it a worthwhile astronomical transient survey instrument. A single Desert Transient Factory (DTF) camera covers 1100 degreesÂ². Several of these can be used to tile the whole sky (12800 degreesÂ²), down to magnitude 13, every 10 seconds. Each patch of sky is going to be monitored by 2 distant cameras continuously. This baseline strategy allows easily identification of local phenomena (sensor artefacts, satellites), and gives weather robustness to the system. The Desert Transient Factory will generate 6TB of data per night. Most of the processing cannot be done in real-time because of limited computing power (solar power), and images cannot be downloaded directly to a datastore because of low bandwidth communications. However, some particular transients needing rapid follow-up (eg. supernovae) can be processed, rapidly corroborated with the mirror system, to trigger follow-up alerts on the most energetic phenomena happening in the universe.

---

**O9.4 Sean Carey**
Spitzer Science Center / IPAC

*Final Calibration and Processing of Warm IRAC Data*

The Spitzer Space Telescope has been conducting a wide range of science investigations including measurement of atmospheric properties of exoplanets and masses of the most distant galaxies during the post-cryogenic operations phase which started in 2009. These investigations using the Infrared Array Camera (IRAC) at 3.6 and 4.5 um will likely continue through 2018 when the James Webb Space Telescope will succeed Spitzer. In preparation for the eventual end of the mission and exploiting the excellent stability of the instrument and spacecraft, we have finalized the data pipeline and most of the calibrations for the IRAC instrument in advance of the mission end to minimize the cost of the closeout process. I present the key modifications made as part of the final pipeline development. The calibrations for the warm mission phase have been substantially revised with the absolute photometric calibration performed with the same methodology as the final cryogenic calibration. Updates to the processing methods due to the longevity of the mission will be highlighted and measurements of the stability of the instrument and resulting data will be discussed.

---

## SESSION X: VISUALISATION AND INNOVATIVE USER INTERFACES

**O10.1 Xiuqin Wu**
California Institute of Technology

*IPAC Firefly Development Roadmap*

*Authors: Xiuqin Wu, David Ciardi, Gregory Dubois-Felsmann, Tatiana Goldina, Steve Groom, Loi Ly, Trey Roby*

IPAC Firefly package has been developed in IRSA (NASA/IPAC Infrared Science Archive) in last six years. It is a software package utilizing state-of-the art AJAX technology to provide an interactive web user interface for astronomers. It has been used to build Spitzer Heritage Archive, WISE Image Service, Planck Visualization, PTF Image Service, and the new IRSA finder chart. It provides three major components: table display, FITS images visualization, and 2D plot. All three highly interactive components can work together using the same data model or separately to provide any combinations of interactivities among them. Firefly toolkits provide an easy way to put interactivities in an otherwise static web page. With a few lines of simple JavaScript embebded in a web page, Firefly toolkits can add manipulative functions to a static table, display a FITS image, or draw an XY 2D plot interactively. At 2015 AAS, we announced that Firefly will be open source under BSD 3-clause license. It is now available on GitHub. IPAC is now responsible for development of LSST Science User Interface/Tools(SUI/T). To satisfy LSST user requirements and to give users more flexibility to build a customized user interface to their specific science needs, we plan to extend Firefly - add more visualization functionalities; make its components more independently accessible; enable users to display images, tables, and 2D XY plots within iPython notebook; allow Firefly tools to be controlled by Python script. In this talk, we will outline the development roadmap with detailed functions and features in next three years.

---

**O10.2 Pragya Mohan**
Victoria University of Wellington

*Three Tools to Aid Visualisation of FITS Files for Astronomy*

Increasingly there is a need to develop astronomical visualisation and manipulations tools which allow viewers to interact with displayed data directly, in real time and across a range of platforms. In addition, increases in dynamic range available for astronomical images with next generation telescopes have led to a desire to develop enhanced visualisations capable of presenting information across a wide range of intensities. This paper describes three new tools for astronomical visualisation and image manipulation that are the result of a collaboration between software engineers and radio astronomers. The first tool aids the visualisation and manipulation of 2D fits images. The tool supports the interactive creation of free-form masks which allow the user to extract any (potentially non-contiguous) subset of a fits image. It also supports annotations which can be placed without affecting the underlying data. The second tool is a fast interactive 3D data cube viewer designed to allow real-time interactive comparisons of multiple spectral line data cubes simultaneously. The final tool is an R package for applying high dynamic range compression

techniques to 2D fits images. This allows the full range of pixel brightness to be imaged in a single image, simultaneously showing the detail in bright sources while preserving the distinction of faint sources. Here we will present these three tools and demonstrate their capability using images from a range of Australian-based radio telescopes.

---

**O10.3**

**Adam Gauci**
University of Malta

*Hybrid, Multi-Frame and Blind Astronomical Image Deconvolution Through L1 and L2 Minimisation*

The study of images in scientific fields such as remote sensing, medical imaging and astronomy comes natural not only because pictures simulate one of the main sensory elements of humans, but also because they allow for the visualisation of wavelengths to which the eyes are not sensitive. However, accurate information extraction from images can only be realised if the data is known to be noise free, blur free and that it contains no artificial artefacts. In astronomical images, apart from hardware limitations, biases are introduced by phenomena beyond control such as for instance atmospheric and ionospheric degradations. The resulting combined blur function is not constant in time nor space and vary according to turbulence in the air column as well as the wavelengths being recorded.

The deconvolution process attempts to recover the true values from the measured intensities. Having a robust and accurate deconvolution algorithm is very important especially for mega-dimensional telescopes such as the Square Kilometre Array (SKA) through which sensitive investigations including gravitational lensing research and the detection of faint sources, are to be made. Despite the non-uniqueness and noise sensitivity, a lot of research in deconvolution methods have been carried out by major scientific committees including those focusing on computer vision, machine learning, optics and astronomy.
Most of the available techniques assume the blur filter to be known and attempt recovery by using this information. While the PSF of the instrument may be resolved with very high accuracy, astronomical images contain random spatially varying factors that change on millisecond scales. For instance in the optical range, the PSF can only be taken to be constant over an isoplanatic patch for 5 to 15ms across regions between 10 and 20cm. Longer exposure times will cause the high frequency components to average out while shorter recording times will not be enough to record all information. In such cases, the true blur kernel cannot be accurately known and blind deconvolution methods have to be used.

Finding an inverse solution that estimates the original scene is an ill-posed problem. Iterative methods are normally applied to try and improve the quality by repetitively apply a task until some predefined stopping criteria are met. Other algorithms are based on Bayesian methods or attempt to enhance images in different basis domains. Although the best results are obtained through the use of specifically designed algorithms that work on signals with a particular set of properties, most techniques still focus on finding a generic model that can be universally applied.

In this work, we investigate the improvements gained if a number of algorithms are used to minimise the overall recovery error. A hybrid image deblurring estimator that processes multiple frames to improve on a global reconstruction, is presented. Most of the available similar methods assume a batch mode of operation and require the entire set of frames to be given prior to the initialising of the recovery process. The presented technique works on each frame individually and hence avoids the need for large memory requirements. For every given image, the Point Spread Function (PSF) is first estimated by minimisation of the L1 norm or L2 norm residuals. A similar search is then carried out to deblur the image using the estimated PSF. Blurred datasets are generated through the combination of Gaussian and Kolmogorov filters. The degradation in performance when noisy images are used is also investigated. Quantification of the accuracy is achieved through the Mean Square Error (MSE). The results from the preliminarily implemented prototype are very satisfactory and encourage further research.

---

**O10.4**

**Christopher Fluke**
Swinburne University of Technology

*The Ultimate Display*

Astronomical images and datasets are increasingly high-resolution and multi-dimensional. The vast majority of astronomers perform all of their visualisation and analysis tasks on low-resolution, two-dimensional desktop monitors. If there were no technological barriers to designing the ultimate immersive stereoscopic display for astronomy, what would it look like? What capabilities would we require of our compute hardware to drive it? And are existing technologies even close to providing a true 3D experience that is compatible with the depth resolution of human stereoscopic vision? With the CAVE2 (an 80 Megapixel, hybrid 2D and 3D virtual reality environment directly integrated with a 100 Tflop/s GPU-powered supercomputer) and the Oculus Rift (a low-cost, head-mounted display) as examples at opposite financial ends of the immersive display spectrum, I will discuss the changing face of high-resolution, immersive visualisation for astronomy.

---

**O10.5 Faviola Molina**

University of Chile

*AstroCloud: An Agile Visualization Platform for Specific Analyses of Astronomical Images*

Visualizing astronomical data is notoriously resources consuming. For example, current visualization tools require a depth knowledge of the source code to let a practitioner to customize them in order to make specific analysis. Moreover, the visualization and navigation through a 3D dataset is usually made as slices of the data. Although there are some 3D platforms for volume rendering, they are limited to show different levels as color coded surfaces. We developed a 3D rendering application, which is able to display the values of the dataset through a color bar. In addition, the color bar can

be manipulated in order to render a subrange of the dataset. This subrange is set by the user's choice. This paper reports on visualizing a 512^3 pixels datacube using a GPU-rendered raycasting.

---

**O10.6 Guido De Marchi**
European Space Agency

*Promoting Scientific Creativity in Astronomy*

The ESAC Science Data Centre (ESDC) provides services and tools to access and retrieve observations and data from all ESA space science missions (astronomy, planetary, and solar-heliospheric). We have recently developed a new suite of user-friendly web-based applications that are easy to use and allow the seamless exploitation of the scientific data from current and past ESA astrophysics missions. In this talk I will touch on the rationale behind an approach that aims to stimulate scientific curiosity and creativity in the astronomical community by making it easier to literally see the same scene in a different light. I will illustrate some of the new services offered by the ESDC, in particular the European HST Archive and the Astronomy Multi Mission Interface, which provide full access to the entire sky as observed with ESA missions.

---

## SESSION XI: KNOWLEDGE DISCOVERY AND DATA MANAGEMENT TOOLS FOR ASTRONOMICAL BIG DATA

**O11.1 Sarah Kendrew**
*Oxford University*

*7 years of .Astronomy: Building the astronomy community of the future*

Technological innovation has changed the face of scientific research: the growth of the web and mobile computing, the cost reduction of data storage and processing power, and advances in large format detector arrays have brought us to an era of "always on" connectivity and big data. The .Astronomy conferences, now in their 7th year, bring together a diverse community of scientists, developers, engineers and educators to discuss innovative ways of exploiting this paradigm for research and public engagement in an open conference format, led by the participants themselves. The events are an opportunity for a young generation of scientists with exceptional computing or creative skills to share their knowledge and ideas, and collaborate outside of the confines of their day to day research. This philosophy is inspiring other conference organisers, and Astronomy Hack Days are now routinely organised at AAS meetings, the UK NAM, and SPIE conferences. Our aim is for .Astronomy to be!

A driving force in building a creative, dynamic and innovative community of astronomers. In this talk, I will show some noted successes and lessons learnt from .Astronomy, describe how the conference has grown and provide perspectives on its future.

---

**O11.2 Mark Allen**
Centre de Donnes astronomiques de Strasbourg

*A Hierarchical Approach to Big Data*

The increasing volumes of astronomical data require practical methods for data exploration, access and visualisation. The Hierarchical Progressive Survey (HiPS) is a HEALPix based scheme that enables a multi-resolution approach to astronomy data from the individual pixels up to the whole sky. We highlight the decisions and approaches that have been taken to make this scheme a practical solution for managing large volumes of heterogeneous data. Early implementors of this system have formed a network of HiPS nodes, with some 230 diverse data sets currently available, with multiple mirror implementations for important data sets. We show how this hierarchical approach can be adapted to expose Big Data in different ways, such as for visual and statistical summaries of large data sets including the CADC HST image archive. Tests with ALMA data cubes show that the calculation and usability of individual 5TB HiPS data sets is feasible. We describe how the ease of implementation, and local customisation of the Aladin Lite embeddable HiPS visualiser have been keys for promoting collaboration on HiPS and the testing of new possibilities.

---

**O11.3 Christophe Arviset**
ESA-ESAC

*The VO: A Powerful tool for global astronomy*

*List of authors: Arviset, Christophe, ESAC Science Data Centre; Allen, Marc, CDS; Aloisi, Alessandra, STScI / MAST; Berriman, Bruce, IPAC; Boisson, Catherine, Observatoire de Paris / CTA; Cecconi, Baptiste, Observatoire de Paris / Europlanet/VESPA; Ciardi, David, NASA Exoplanet Science Institute, IPAC; Evans, Janet, SAO / CXC; Fabbiano, Giuseppina, SAO / CXC; Genova, Francoise, CDS; Groom, Steve, IRSA / IPAC; Jenness, Tim, LSST; Mann, Bob, Institute for Astronomy, University of Edinburgh / WFCAM, VISTA; McGlynn, Tom, NASA / HEASARC; O'Mullane, Wil, ESA-ESAC / Gaia; Schade, David, CADC; Stoehr, Felix, ESO / ALMA; Zacchei, Andrea, INAF-OATs / Euclid;*

Since its inception in the early 2000's, the Virtual Observatory has become a major factor in the discovery and dissemination of astronomical information worldwide. It has been developed as a collaboration of many national and international projects. The International Virtual Observatory Alliance (IVOA) has been coordinating all these efforts worldwide to ensure a common VO framework that enables transparent accessibility and interoperability to astronomy resources (data and software) around the world.

The VO is not a magic solution to all astronomy data management challenges but it does bring useful solutions to many. VO interfaces are broadly found in astronomy's major data centres and projects worldwide. Astronomy data centres have been building VO services on top of their existing data

services to increase interoperability with other "VO-compliant" data resources. The continuous and increasing development of VO applications (ie Aladin, Topcat, Iris) greatly facilitates multi-instruments, multi-wavelengths science.

More recently, several major new astronomy projects are directly adopting VO standards to build their data management infrastructure, giving birth to "VO built-in" archives (eg CADC, Gaia, CTA,). Embracing VO framework from the beginning brings the double gain of not needing to reinvent the wheel and ensuring from the start interoperability with other astronomy VO resources. Some of the IVOA standards are also starting to be used by neighbour disciplines like planetary sciences.

There is still quite a lot to be done on the VO, in particular tackling the upcoming big data challenge and how to find interoperable solutions to the new data analysis paradigm of bringing and running the software close to the data.

While we report on the current status, this paper also expresses the desire of the presenting astronomy data centres and projects who are developing and adopting the VO technology to engage others to join the effort, to ensure that the VO standards fits new astronomy projects requirements and needs!

---

**O11.4 Slava Kitaeff**
UWA

*VO Services with JPEG2000 Client-Server Visualisation: Astronomy Data Services at Pawsey Supercomputing Centre*
*Authors: S.Kitaeff, D. Marrable, J.T. Mararecki, A. Wicenec, C. Wu, J. Harrison*

There is an immense, internationally significant collection of radio astronomy data in Australia, generated by organisations such as CSIRO and ICRAR, which are also plying an active role in building the Square Kilometre Array (SKA). Australia has constructed two of the three official SKA pathfinders: the Australian SKA Pathfinder (ASKAP) and the Murchison Widefield Array (MWA), so the collection of data will grow in the near future. Commonwealth (Super Science) has made a considerable infrastructure investment to support Data Intensive Sciences within the Pawsey Supercomputing Centre, MWA and ASKAP. The scientists use the co-located high performance compute and data stores to facilitate the research. Research Data Service (RDS) is an investment to support Data Intensive Sciences, such as e.g. MWA GLEAM survey, by providing an infrastructure to store large datasets. RDS already hosts many PBs of MWA data.

Astronomy Data Services (ADS) project has developed a solution to provide public access to astronomy data stored on RDS infrastructure. Together with IVOA services, such as TAP, SIAP and ADQL, JPEG2000 encoding for imagery data, and the consecutive streaming client-server visualisation using JPIP protocol have been enabled. FITS imagery data gets encoded upon SIAP request by the user. Encoded image or image-cube is then made available either through download or streaming services. We have also developed JPEG2000 enabled version of Aladin software that dynamically and progressively streams only the required for visualisation part of the data from JPIP server at requested quality/resolution. Ingest of data from the local storages is enabled by NGAS.

**O11.5 Jessica Chapman and James Dempsey**
CSIRO Astronomy and Space Science

*The CSIRO ASKAP Science Data Archive:*

ASKAP is an array of 36 radio antennas, located at the Murchison Radio Observatory, in Western Australia. The ASKAP antennas are equipped with innovative phased array feed receivers. These provide an extremely wide field-of-view and enable ASKAP to carry out sensitive large-scale surveys of the Southern Sky. ASKAP data volumes are very high. In full operations, the CSIRO ASKAP Science Data Archive (CASDA) will archive and manage around 5 PetaBytes of data each year.

Radio astronomy signals collected with the ASKAP antennas are transferred to the Pawsey Supercomputing Centre in Perth where they are processed, archived, and made available to astronomers. Astronomers will interact with ASKAP data products through the CASDA application. ASKAP data collections from the major surveys will be openly available to the global astronomy community. CASDA will provide search and discovery tools using the CSIRO Data Access Portal (DAP) and international Virtual Observatory (VO) protocols. The first CASDA production release is due for November 2015, prior to the start of Early Science in early 2016. In this talk we will describe the CASDA application design and architecture and will demonstrate how it can be used for data discovery and access through both web-based services and VO tools.

---

## SESSION XII: DATA PIPELINES

**O12.1 Jamie Kinney**
Amazon

*Astronomy Data Pipelines in the Cloud: Serverless Computing with AWS*

Astronomers from around the world are using AWS for a range of applications including data ingestion, processing, archival, analytics and visualization. With the advent of higher-level services such as Lambda, Kinesis, and the Elastic Container Service, it is now possible to create efficient and highly-scalable Astronomy data pipelines without having to administer individual servers. Jamie Kinney, AWS Direct of Scientific Computing (a.k.a. “SciCo”) will describe the many ways that AWS is helping researchers increase their pace of scientific discovery and will explore these newer technologies in more detail. This session will also describe recent advancements in Amazon’s networking, HPC and data management and analytics capabilities as they relate to Scientific Computing.

---

**O12.2 Daniel Durand**
National Research Council Canada

*HST in the Clouds*

The HST archive system at CADC has been evolving constantly since its inception in 1990. After basic upgrades to the associate storage system (optical disks, CDs, DVDs, magnetic disks) and implementing multiple processing system (On the Fly calibration, CACHE), the HST system at CADC is now running in a cloud system (CANFAR). After multiple hurdles mostly caused by the way the HST calibration systems has been designed many years ago, we are now happy to report that the system is running quite nicely under the CANFAR system designed by CADC and operated by compute Canada consortium. Although not very large, the HST collection needs potentially constant recalibration to take advantage of new software and re calibration files. This talk describes the unique challenges in bringing legacy pipeline software to run in a massive cloud computing system. In the cache mode, and using a cloud system, the HST processing is now running in a system which could be easily scaled. Presently more than 200 cores could be used to process the HST images, and this could potentially grown to 1000s of cores, allowing a very uniform calibrated archive since any perturbation to the system could be dealt with in a few hours.

---

**O12.3 Howard Bushouse**
Space Telescope Science Institute

*The JWST Data Calibration Pipeline*

The James Webb Space Telescope (JWST) is the successor to the Hubble Space Telescope (HST) and is currently expected to be launched in late 2018. The Space Telescope Science Institute (STScI) is developing the pipeline systems that will be used to provide routine calibration of the science data received from JWST. The JWST calibration pipelines use a processing environment provided by a Python module called "stpipe" that provides many common services to each calibration step, relieving step developers from having to implement such functionality. The stpipe module provides multi-level logging, command-line option handling, parameter validation and persistence, and I/O management. Individual steps are written as Python classes that can be invoked individually from within Python or from the stpipe command line. Pipelines are created as a set of step classes, with stpipe handling the flow of data between steps. The stpipe environment includes the use of standard data models. The data models, defined using json schema, provide a means of validating the correct format of the data files presented to the pipeline, as well as presenting an abstract interface to isolate the calibration steps from details of how the data are stored on disk.

---

**O12.4 Cormac Purcell**
The University of Sydney

*The POSSUM Pipeline: Getting Results from ASKAP Early Science*

The Polarisation Sky Survey of the Universe's Magnetism (POSSUM) project will measure the Faraday rotation of over three million distant galaxies, dramatically improving our understanding of astrophysical magnetism. Early science observations on the Australian Square Kilometre Array Pathfinder (ASKAP) will begin in January 2016, eventually producing over 3TB of data per day. Verified results have been promised to the community within days of being observed, posing significant challenges to analyse the data in a timely fashion. Moreover, polarisation observations are multi-dimensional and require complex analysis techniques to tease out reliable scientific results. Here we present a prototype POSSUM analysis pipeline intended to facilitate quality control during observations and to create a scientifically excellent catalogue soon after the data have been taken. The software makes use of techniques such as rotation measure synthesis, advanced model fitting and Bayesian model comparison to characterise source polarisation properties. Pipeline users can visualise the results using a custom graphical interface written in Python, which presents options to run defined queries or perform custom analysis. We have tried to ensure that the pipeline is robust enough to be used with data from many telescopes and offer it as a resource under an open licence.

---

**O12.5 Andreas Wicenec**
University of Western Australia

*DROP Computing: Data Driven Pipeline Processing for the SKA*

The correlator output of the SKA arrays will be of the order of 1 TB/s. That data rate will have to be processed by the Science Data Processor using dedicated HPC infrastructure in both Australia and South Africa. Radio astronomical processing in principle is thought to be highly data parallel, with little to no communication required between individual tasks. Together with the ever increasing number of cores (CPUs) and stream processors (GPUs) this led us to step back and think about the traditional pipeline and task driven approach on a more fundamental level. We have thus started to look into dataflow representations [1] and data flow programming models [2] as well as data flow languages [3] and scheduling [4]. We have investigated a number of existing systems and prototyped some implementations using simplified, but real radio astronomy workflows. Despite the fact that many of these approaches are already focussing on data and dataflow as the most critical component, we still missed a rigorously data driven approach, where the data itself is essentially driving the whole process. In this talk we will present the new concept of DROP Computing (condensed data cloud), which is an integral part of the current SKA Data Layer architecture. In short a DROP is an abstract class, instances of which represent data (DataDrop), collections of DROPs (ContainerDrop), but also applications (ApplicationDrop, e.g. pipeline components). The rest are 'just details', which will be presented in the talk.

*[1] Jack B. Dennis, David P. Misunas. A Preliminary Architecture for a Basic Data-Flow Processor, MIT, IN PROCEEDINGS OF THE 2ND ANNUAL SYMPOSIUM ON COMPUTER ARCHITECTURE, 1975; [2] Alan L. Davis. Data driven nets: A maximally concurrent, procedural, parallel process representation for distributed control systems. Technical report, Technical Report, Department of Computer Science, University of Utah, Salt Lake City, Utah, 1978; [3] W. M. Johnston, J. R. P. Hanna, and R. J. Millar, "Advances in dataflow programming languages," ACM Comput. Surv., vol. 36, no. 1, pp. 1–34, 2004.; [4] A. Benoit, U. Catalyurek, Y. Robert, E. Saule, A. Benoit, U. Catalyurek, Y. Robert, E. Saule, A. Survey, and P. Workflow, "A Survey of Pipelined Workflow Scheduling : Models and Algorithms," HAL open Arch., 2014.*

---

**O12.6 Kevin Vinsen**
ICRAR

*Imaging SKA-Scale Data in Three Different Computing Environments – the Cloud, a Supercomputer, and an In-House Cluster*

CHILES, the Cosmic HI Large Extragalactic Survey, is a very sensitive search for extra galactic emission from the Jansky Very Large Array, managed by the National Radio Astronomy Observatory. The upgraded J-VLA is capable of producing prodigious data volumes, which overwhelms most approaches to the data flow management. The problem is similar in scale to that associated with the SKA.

We have investigated how one would manage to process these data volumes using three very different computing environments: a moderate sized cluster, such as a group like ICRAR could (and does) host and control; a high performance computing cluster that would be provided by a national facility such as the Pawsey centre and a cloud computing environment, such as provided by the Amazon Web Service (AWS). This allowed us to explore three very different approaches, all of which would be of the scale accessible to groups such as ours via in-house capital expenditure, via competitive applications for resources on national infrastructure or via cumulative operational expenditure.

We report on the advantages and disadvantages of all of these environments and draw conclusions as to what the most important issues are in delivering SKA-scale science.

---

**O12.7 Christopher Hollitt**
Victoria University of Wellington

*An Overivew of the SKA Science Analysis Pipeline*

When completed the Square Kilometre Array (SKA) will feature an unprecedented rate of image generation. While previous generations of telescopes have relied on human expertise to extract scientifically interesting information from the images, the sheer data volume of the data will now make this impractical. Additionally, the rate at which data are accrued will not allow traditional imaging products to be stored indefinitely for later inspection meaning there is a strong imperative

to discard uninteresting data in pseudo-real time. Here we outline components of the SKA science analysis pipeline being developed to produce a series of data products including continuum images, spectral cubes and Faraday depth spectral. We discuss a scheme to automatically extract value from these products and discard scientifically uninteresting data. This pipeline is thus expected to give both an increase in scientific productivity, and offers the possibility of reduced data archive size producing a considerable saving.

---

## SESSION XIII: KNOWLEDGE DISCOVERY AND DATA MANAGEMENT TOOLS FOR ASTRONOMICAL BIG DATA

### O13.1 Bradley Whitmore

Space Telescope Science Institute

*Version 1 of the Hubble Source Catalog*

*Whitmore, B., Budavari, T., Donaldson, T., Downes, R., Lubow, S.,Quick, L., Strolger, L., Wallace, G., White, R. L.*

The Hubble Source Catalog (HSC) is designed to help optimize science from the Hubble Space Telescope by combining the tens of thousands of visit-based Hubble Legacy Archive (HLA - available at http://hla.stsci.edu) source lists into a single master catalog. The HSC includes ACS/WFC, WFPC2, and WFC3 source lists generated using the Source Extractor software (Bertin & Arnouts 1996). The current version of the catalog includes roughly 80 million detections of 30 million objects involving 112 different detector/filter combinations and about 50 thousand HST exposures cross-matched using techniques described in Budavari & Lubow (2012). To carry out the cross matching we first improve the astrometry for the source lists using a histogram method to compare against Pan-STARRS, SDSS, and 2MASS catalogs. We then further improve the alignment using a highly efficient method for approximately aligned source lists to achieve relative offsets typically good to a few milli-arcsec. A final pass versus Pan-STARRS is use to normalize to their absolute astrometry grid. The astrometric residuals for HSC objects are typically within 10 mas and the magnitude residuals between repeat measurements are generally within 0.10 mag. The primary ways to access the HSC are the MAST Discovery Portal (http://mast.stsci.edu), and a CasJobs capability for advanced searches. The URL for the HSC is http://archive.stsci.edu/hst/hsc/.

---

### O13.2 Bruce Berriman

IPAC, Caltech

*The Next Generation of the Montage Image Mosaic Toolkit.*

The scientific computing landscape has evolved dramatically in the past few years, with new schemes for organizing and storing data that reflect the growth in size and complexity of astronomical data sets. In response to this changing landscape, we are, over the next two years,

deploying the next generation of the Montage toolkit ([ascl:1010.036]). The first release (September 2015) will support multidimensional data sets ("data cubes) and insertion of XMP/AVM tags that allows images to "drop-in"" to the WWT (see this example for M51; http://www.worldwidetelescope.org/wwtweb/ShowImage.aspx?scale=0.4&rotation=180.00&ra=202.48417&dec=47.23056&y=1800.5&x=1801.0&thumb=http://exoplanetarchive.ipac.caltech.edu/workspace/AVM/M51_SDSS_small.jpg&imageurl=http://exoplanetarchive.ipac.caltech.edu/workspace/AVM/M51.jpg&name=M51) The same release will offer a beta-version of an interactive image visualizer which can be used as an application as a web service or in a Python environment. Subsequent releases will support HEALPix (now standard in cosmic background experiments); incorporation of Montage into package managers (which enable automated management of software builds) and support for a library that will enable Montage to be called directly from Python. This next generation toolkit will inherit the architectural benefits of the current engine - component based tools ANSI-C portability across Unix platforms and scalability for distributed processing. With the expanded functionality under development Montage can be viewed not simply as a mosaic engine but as a scalable portable toolkit for managing organizing and processing images at scale. The architectural benefits of Montage provide considerable flexibility to the end user and we will describe how the community is taking advantage of it to integrate its components into pipelines and workflow environments. Examples include: underpinning a pipeline to create three color SDSS mosaics for galaxies in the RC3 catalogs (Lee 2014; [ascl:1411.006]); integration into the AAO/UKST SuperCOSMOS Halpha Survey flux calibration pipeline (Frew et l. 204 MNRAS 440 1080); integration into the processing environment of the Sydney-AAO Multi-object Integral (SAMI) field spectrograph pilot survey (Fogarty et al. 2014; MNRAS 443 485); and integration into the processing environment for the Palomar Transient Factory (Surace et al. 2014; PASP 126 674). In addition it is an exemplar tool for the development of cyberinfrastructure systems that will enable non-experts to run workflows at scale. One example is building AstroTaverna workflows with Virtual Observatory services (Ruiz et al. 2014; Astronomy and Computing 7.3). Another is the production in collaboration with ISI/USC and Amazon Web Services of a 16-wavelength Atlas of the Galactic Plane with Open Source tools such as the Pegasus Workflow management system which when complete is aimed at deploying a set of tools for scientists to process and manage data on distributed platforms (Berriman et al. 2015; http://www.noao.edu/meetings/bigdata/files/Berriman.pdf). Montage is funded by the National Science Foundation under Grant Number ACI-1440620.

---

**O13.3 Santhilata Kuppili Venkata**

King's College London

*Adaptive Caching Using Sub-Query Fragmentation for Reduction in Data Transfers from Distributed Databases*

One of the challenges with Big Data is to transfer massive amounts of data from data server(s) to users. Unless data transfers are planned, organized and regulated carefully, they can become a potential bottleneck and may lead to change the way of querying databases and even the design of the backend data structures. This is a pronounced problem in the case of virtual observatories where

data is to be brought from multiple astronomical databases all around the world. To reduce data transfers here we propose an adaptive middleware caching using sub-query caching technique.

Sub-query caching technique involves fragmenting the query into smaller sub queries. A sub-query is defined as the part of the query which can be separated as a block of data or a data object. This technique applies association rules over the database-specific data localization during the query processing to identify the optimum grain of data to be cached. As the query is cached as smaller objects, it achieves reduction in the processing costs needed for joins. Also reduction in the data transfers is achieved as parts of the query is already present at the cache.

A distributed database environment is simulated incorporating the key features of real life scenario with multiple user groups querying through common query interface. Synthetic query sets are generated for input with varied complexity and sub-query repetition. Initial experiments performed with these input sets showed considerable reduction in the response time when used our approach compared to full query cache method. We used association algorithms and decision trees for cache training and maintenance. Experiments showed reductions in data transfers needed with our fully trained cache compared to the amount of data transfers needed when entire columns of data to be transferred from data server to the middleware location.

Future work includes (i) developing a mobile architecture with central control for cache units based on the popularity of the data generated from data usage patterns and (ii) query approximation to estimate the exact need of data.