

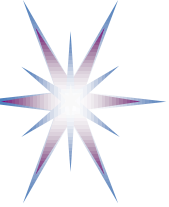
# Big Data Challenges for e-Science Infrastructure



**AAA-Study Project**

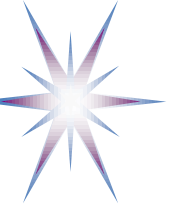
Yuri Demchenko,  
SNE Group, University of Amsterdam

COINFO2012 Conference  
24-25 November 2012, Nanjing, China



# Outline

- Introduction
  - System and Network Engineering (SNE) group at the University of Amsterdam (UvA)
- Big Data Science as the next technology development focus
- European Research Area (ERA) and SDI in Europe
- Data categories and Data Lifecycle in modern e-Science
- General requirements to e-Infrastructure for Big Data Science
- Defining SDI architecture framework
  - Clouds as an infrastructure platform for complex/scientific data
- Questions and Discussion



# Contributing projects (FP7)

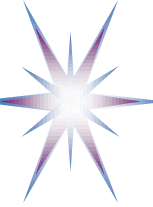
- AAA Study: Study on AAA Platforms For Scientific data/information Resources in Europe -  
<https://confluence.terena.org/display/aaastudy/AAA+Study+Home+Page>
  - Analysis of existing Research Infrastructure in Europe use cases and requirements
  - Final report and recommendations to EC, MS and NREN's published
- GEYSERS – Generalised Architecture for Infrastructure services -  
<http://www.geysers.eu/>
- GEANT3 JRA3 Task 3 – Composable services (GEMBus) -  
<http://www.geant.net/>

## Cloud/InterCloud oriented research

- Intercloud Architecture for Interoperability and Integration, Release 1, Draft Version 0.4. SNE Technical Report 2012-03-02, 19 June 2012  
<http://staff.science.uva.nl/~demch/worksinprogress/sne2012-techreport-12-05-intercloud-architecture-draft04.pdf>
  - (1) Generic Cloud IaaS Architecture, Release 1, 15 April 2011  
Published as <http://staff.science.uva.nl/~demch/worksinprogress/sne2011-techreport-2011-03-clouds-iaas-architecture-release1.pdf>
  - (2) InterCloud OS/Middleware (low level Intercloud integration)
- Security Infrastructure for Cloud (dynamically provisioned)
- Contributing to cloud standardisation by OGF, NIST, IEEE, IETF
- Implementation – EU projects GEYSERS, GEANT3, COMMIT
  - Telco and NREN driven – core network and last mile

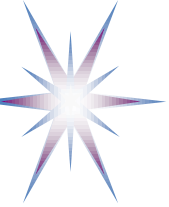
## Challenging infrastructure issues in Big Data technologies

- Distributed data support, high-performance optical networks, trustworthy ICT



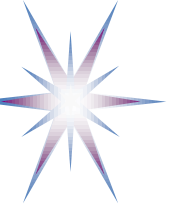
# Big Data Science as the next technology focus

- *Big Data is becoming the next buzz word*
- Based on the e-Science concept and entire information and artifacts digitising
  - Requires also new information and semantic models for information structuring and presentation
  - Requires new research methods using large data sets and data mining
    - Methods to evolve and results to be improved
- Changes the way how the modern research is done (in e-Science)
  - Secondary research, data re-focusing, linking data and publications
- Big Data require **infrastructure** to support both distributed data (collection, storage, processing) and metadata/discovery services
  - Demand for trusted/trustworthy infrastructure
  - Clouds provide just right technology for infrastructure virtualisation to support different data set



# EC Strategy and Policy on STI (Scientific and Technology Information)

- Covers 2 main areas
  - Scientific Information policies defining priorities for STI in Europe and harmonise national policies
  - e-Infrastructure for scientific information
    - Management of scientific data during their whole lifecycle: includes creation, access, (re)use, and preservation
- Main targets
  - New technologies for research and scientific data use
  - Better access to support information exchange and cooperation
  - Preservation of data for future reuse and secondary research
- EC initiative on Open Access scientific publications from publicly funded projects
- G8+O5 Global Research Data Infrastructure, Subgroup on Data, Draft Report, 28 October 2011
  - Requires “reliable infrastructures for persistent identification of data (e.g. digital object identifiers, handle systems), researchers (e.g. digital authors identifiers), and authentication, authorization and accounting systems (AAA)”



# Horizon2020 Consultation Meeting (Rome 11-12 April 2012)

- **Vice-President of the European Commission Mme Neelie Kroes**
  - *"Open e-Infrastructures for Open Science" - The Digital Agenda and Access to Scientific Information*
- **Working Group 1: Open Global Data Infrastructure: towards an international framework for collaborative scientific data infrastructure**
  - Several contributing "position papers" produced about the subject of a "Data Web Forum" or a "Data Access and Interoperability Task Force" and reports such as "Riding the Wave" from the High-Level Group on Scientific Data or the report of the G8+5 working group on Data.
- Working Group 2: Open Scientific Content: e-Infrastructure policies and services to support access, storage, processing and exchange of scientific information
- Working Group 3: Open Research Culture: Engagement of researchers and society with Open Science through collaborative data infrastructure.

## Declaration from Rome meeting

"Researchers and practitioners from any discipline are able to find, access and process the data they need in a timely manner. They are confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.

Data are stored, managed, shared, and preserved in a way that optimizes scientific discovery, innovation, and societal benefit. Where appropriate, producers of data benefit from opening it to broad access and routinely deposit their data in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy."



# Open Access to Scientific Publications

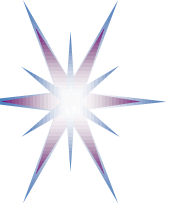
- EC initiative on Open Access scientific publications from publicly funded projects
  - Included into Declaration from Rome meeting
  - Approx 3500 publicly funded ROs and 2000 privately funded ROs
  - Special funding scheme for reimbursing publications
  - Issues with China, India, Russia compliance to OA principles
    - Consultation at high governmental level
- OpenAIRE project exploring models for open access to publications
  - PID (Persistent ID for data), ORCHID (Open Researcher ID), Linked data
- Community initiative - Panton Principles for Open Data in Science (<http://pantonprinciples.org/>)





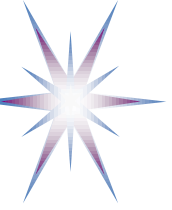
# Coordination in European Research Area (ERA)

- European Commission – *but not only*
- EIROforum – European Intergovernmental Research Organisation
  - Profile committees organised by scientific domain
- ESFRI – European Strategy Forum for Research Infrastructure
  - Coordinates projects and funding for Research Infrastructures (RI) (2002-2010)
- eIRG – e-Infrastructure Reflection Group
  - High level policy development for Europe on e-Infrastructure
- EEF - European e-Infrastructure Forum
  - Principles and practices to create synergies for distributed Infrastructures
- CERN - High Energy Physics and LHC experiment
- TERENA
  - REFEDS – Research and Education Federations
- LIBER – Association of European libraries
  - Growing role of scientific libraries including access to research information



# Big Data Science and European Research Areas (1)

- High Energy Physics (HEP)
  - Running experiment on LHC and infrastructure WLCG (Worldwide LHC Grid)
    - Already producing PBytes of information
    - Worldwide distribution and processing
  - CERN and national HEP centers
- Low Energy Physics and Material Science (photon, proton, laser, spectrometry)
  - Number of research facilities serving international communities
  - Multiple short projects producing TBytes of information
    - Experimental data storage, identification, trusted access to multiple users (including public and private researchers)
- Earth, weather and space observation
  - Climate research and Earth observation
    - With new 4? satellites to be launched starting 2017 to produce PBytes monthly
  - ESA (European Space Agency)



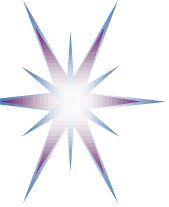
# Big Data Science and European Research Areas (2)

- Life science and biodiversity (Genomic, Biomedical and Healthcare research)
  - Human genome (EMBL-EBI)
    - Currently centralised databases but evolving to distributed
    - ELSI data - Special requirements to data integrity and privacy
  - Living species and biodiversity
    - Mobile/field access, filtering and on-demand computing
    - Public contribution, vocational or citizen researchers
  - Numerous local/offline databases to be brought online
  - Projects: ELIXIR, EGA, LifeWatch
- Humanities (History, languages, human behaviour)
  - Rediscovering research with total information digitising
    - Expected huge amount of data to digitise all human heritage
  - Very spread research community
  - Projects: CLARIN, DARIAH, EUDAT
- Additional: Data collection from sensors and online activities
  - Intelligence, Homeland Security, Log data (+ Facebook data :-)



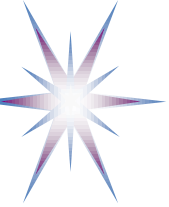
# Existing and emerging Europe wide SDI

- WLCG – Worldwide LHC Grid (CERN, Geneva)
- EGI – European Grid Infrastructure (successor of the EGEE project)
  - Operational Grid infrastructure serving around 10,000 researches worldwide
  - Published “Seeking new horizons: EGI’s role for 2020”
  - Federated Cloud Infrastructure (initiative) provides an infrastructure platform for operational and legacy Grid services
- PRACE – Partnership for Advanced Computing in Europe
- HELIX Nebula – The Science Cloud (prospective cloud based SDI)
  - Private partnership project with wide industry participation but limited EC support
- Growing Research Infrastructures for different research communities
  - CLARIN, EUDAT, LifeWatch, ELIXIR, etc.
    - Less technology and more subject focused



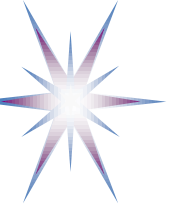
# Big Data Challenges and Initiatives

- Peta and Exa scale problems
  - Storage, Computing, Transfer/Network
  - International Exascale Software Project (<http://www.exascale.org/>)
  - A Vision for Global Research Data Infrastructure (<http://www.grdi2020.eu/>)
- G8+O5 Global Research Data Infrastructure, Subgroup on Data, Draft Report, 28 October 2011
- International Initiative “Research Data Alliance”  
<http://www.rd-alliance.org/>
  - *To accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability.*
  - Consolidates previous initiatives
    - Data Web Forum (DWF) initiative by NSF
    - DAITF – Data Access and Interoperability Task Force initiated by EUDAT project

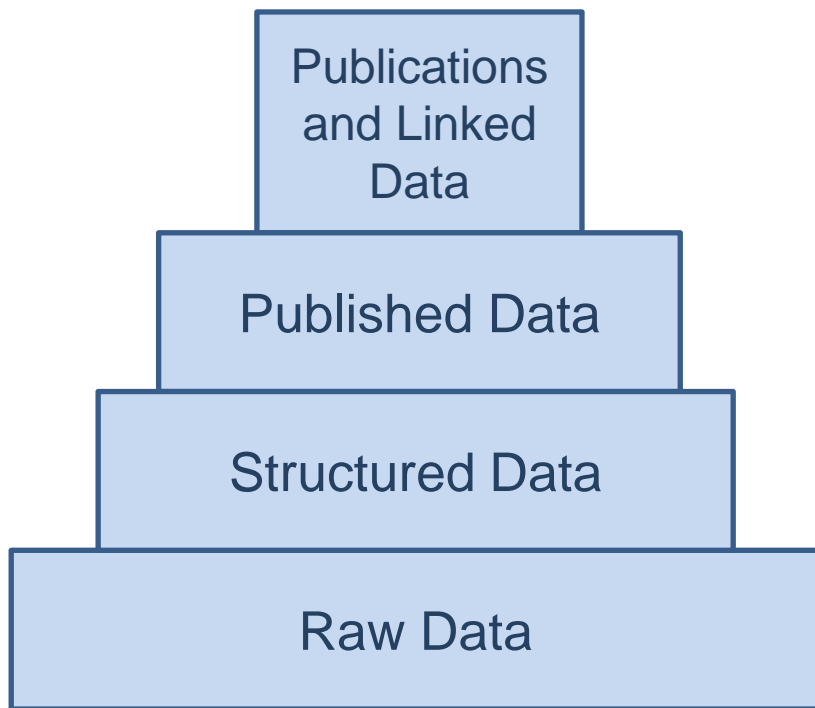


# E-Science Features

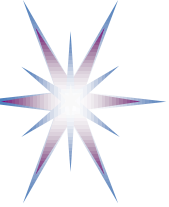
- **Automation** of all e-Science processes including data collection, storing, classification, indexing and other components of the general data curation and provenance
- **Transformation** of all processes, events and products **into digital form** by means of multi-dimensional multi-faceted measurements, monitoring and control; digitising existing artifacts and other content
- Possibility to **re-use** the initial and published research **data** with possible data re-purposing for secondary research
- **Global data availability** and access over the network for cooperative group of researchers, including wide public access to scientific data
- Existence of necessary infrastructure components and management tools that allows fast **infrastructures and services composition, adaptation and provisioning on demand** for specific research projects and tasks
- **Advanced security and access control** technologies that ensure secure operation of the complex research infrastructures and scientific instruments and allow creating **trusted secure environment** for cooperating groups and individual researchers.



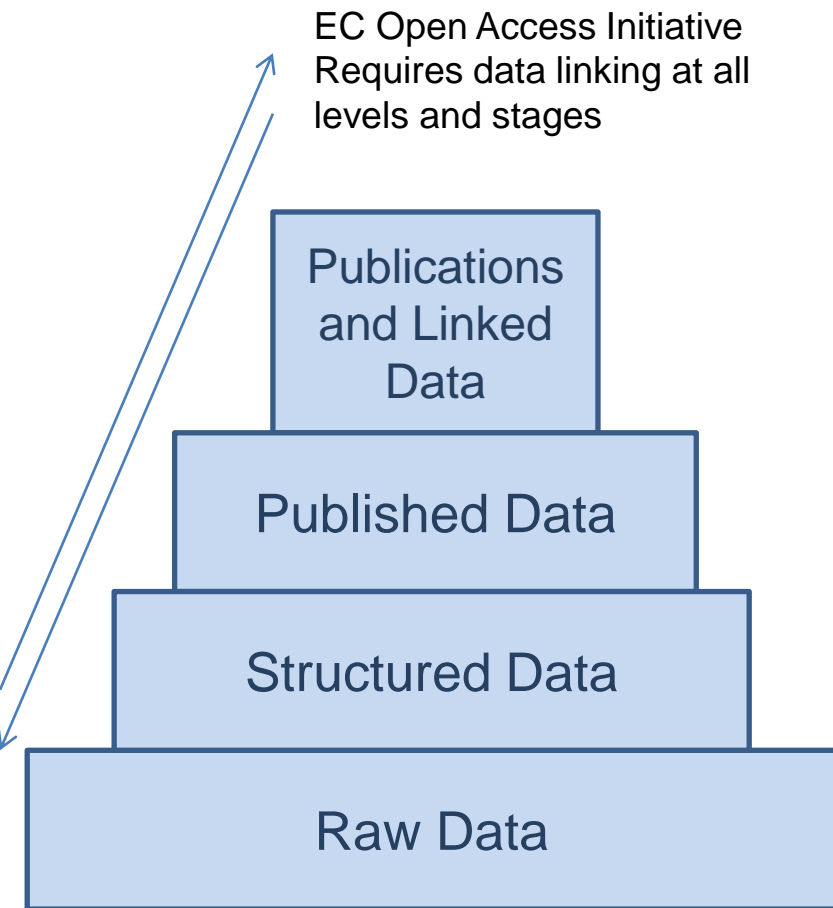
# Scientific Data Types



- **Raw data** collected from observation and from experiment (according to an initial research model)
- **Structured data** and datasets that went through data filtering and processing (supporting some particular formal model)
- **Published data** that supports one or another scientific hypothesis, research result or statement
- **Data linked to publications** to support the wide research consolidation, integration, and openness.

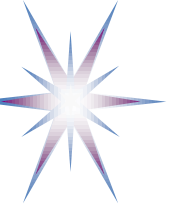


# Scientific Data Types

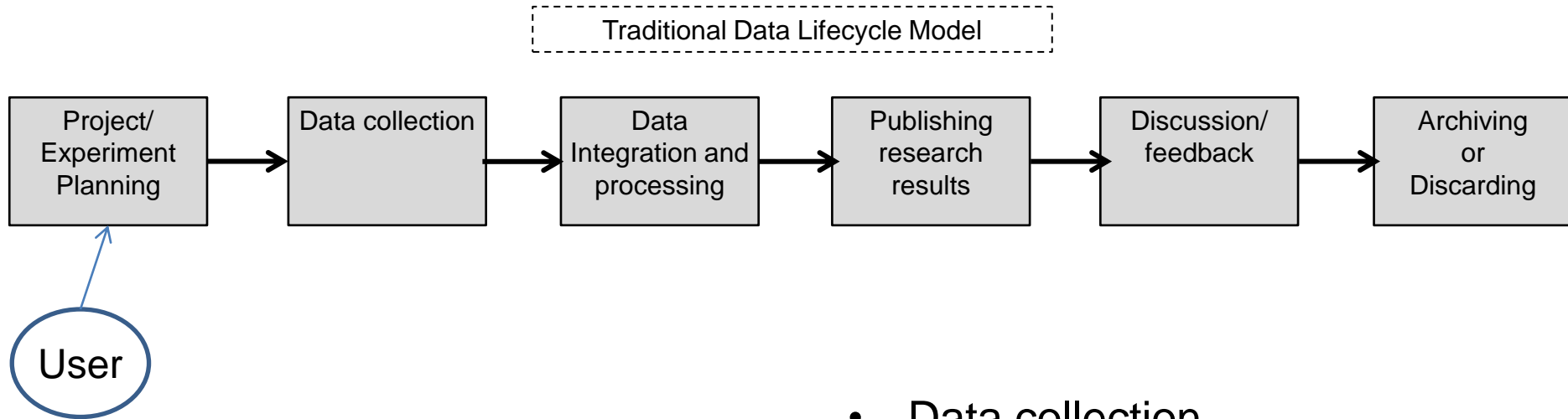


- **Raw data** collected from observation and from experiment (according to an initial research model)
- **Structured data** and datasets that went through data filtering and processing (supporting some particular formal model)
- **Published data** that supports one or another scientific hypothesis, research result or statement
- **Data linked to publications** to support the wide research consolidation, integration, and openness.



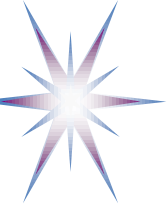


# Traditional Data Lifecycle Model



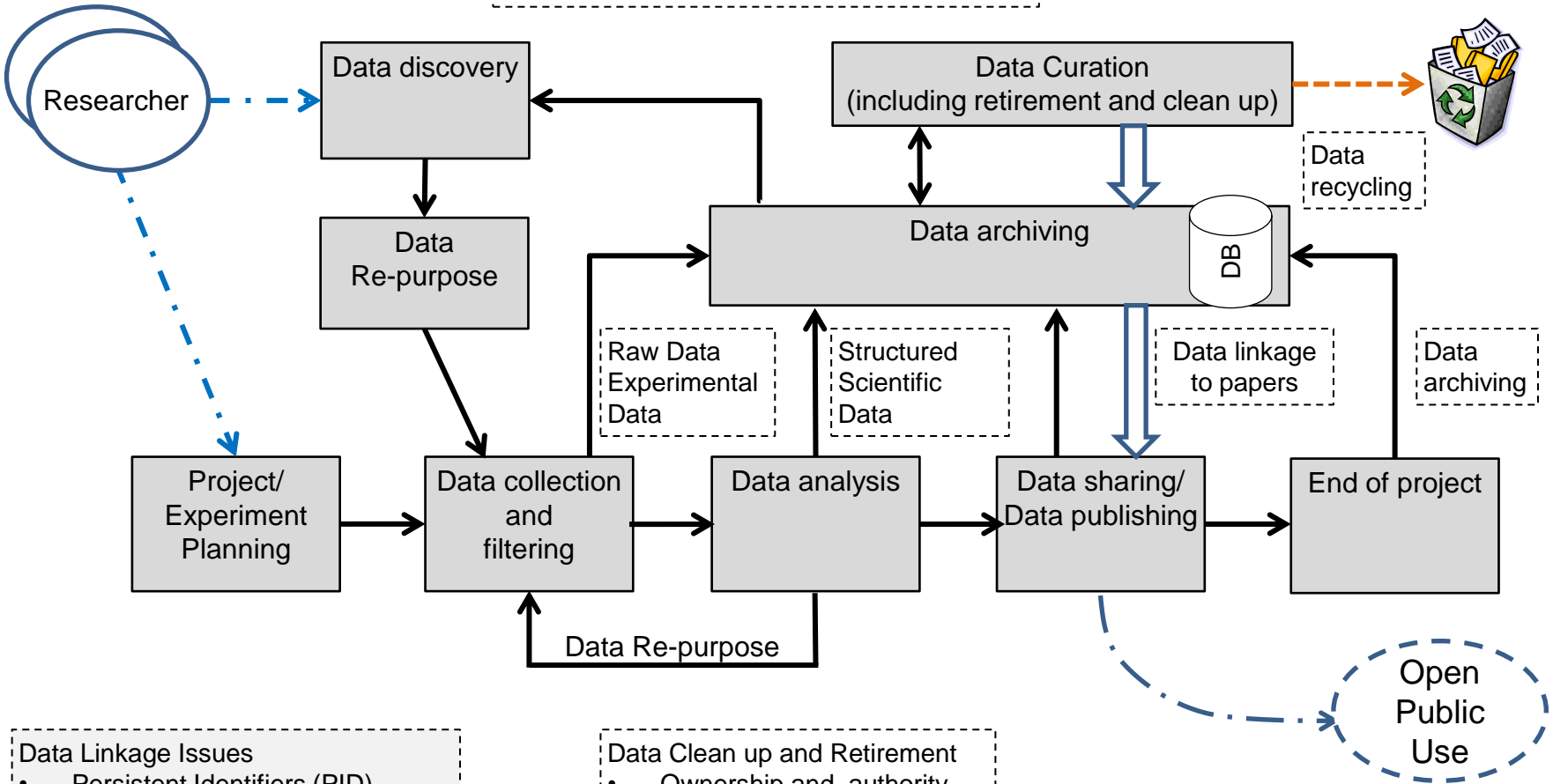
- Data collection
- Data processing
- Publishing research results
- Discussion
- Data and publications archiving

*Lack of initial data preservation and data linking to publications*



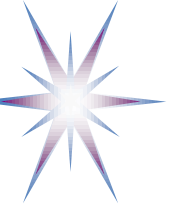
# Data Lifecycle Model in e-Science - II

Data Lifecycle Model in e-Science



- Data Linkage Issues
- Persistent Identifiers (PID)
  - ORCID (Open Researcher and Contributor ID)
  - Lined Data

- Data Clean up and Retirement
- Ownership and authority
  - Data Detainment



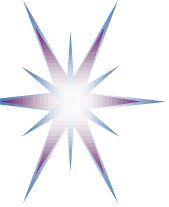
# General requirements to SDI for emerging Big Data Science

- Support for *long running experiments and large data volumes* generated at high speed
- *Multi-tier inter-linked data distribution and replication*
- *On-demand infrastructure provisioning* to support data sets and scientific workflows, mobility of data-centric scientific applications
- Support of *virtual scientists communities*, addressing dynamic user groups creation and management, federated identity management
- Support for the *whole data lifecycle* including metadata and data source linkage
- *Trusted environment* for data storage and processing
- Support for data integrity, confidentiality, accountability
- *Policy binding to data* to protect privacy, confidentiality and IPR

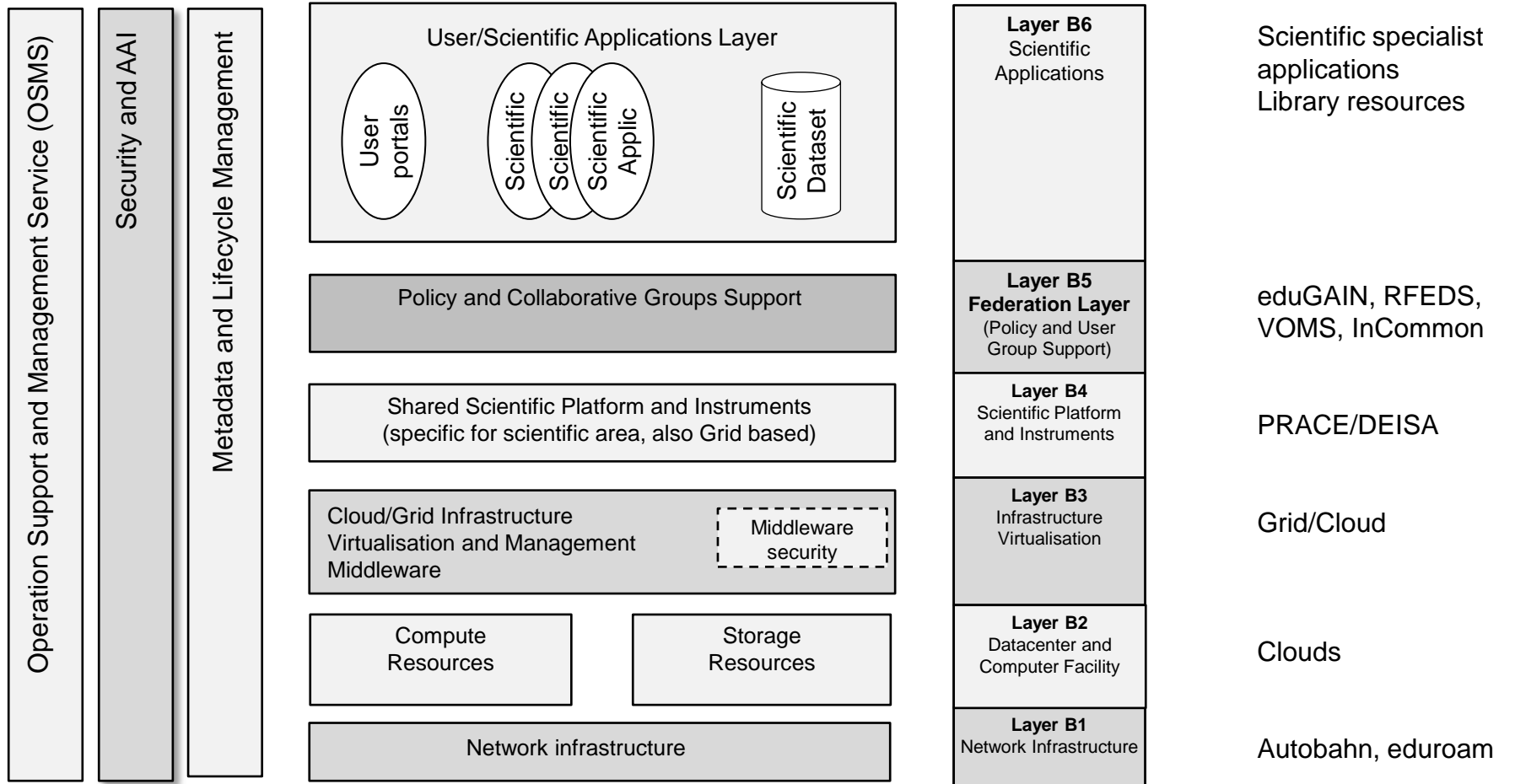


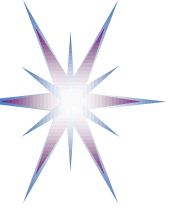
# Defining Architecture framework for SDI and ACAI

- Scientific Data Lifecycle Management (SDLM) model
- e-SDI multi-layer architecture model
- RORA model to define relationship between resources and actors
  - RORA (Resource-Ownership-Role-Actor) model defines relationship between resources, owners, managers, users
  - Initially defined for telecom domain
  - Potentially new actor in SDI – Subject of data (e.g. patient, or scientific object/paper)
- Security and Access Control and Accounting Infrastructure (ACAI)
  - Authentication, Authorisation, Accounting
    - Supported by logging service
  - Extended to support data access control and operations on data
  - Trust management infrastructure



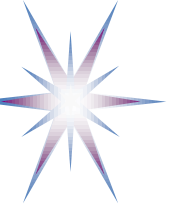
# SDI Architecture Model





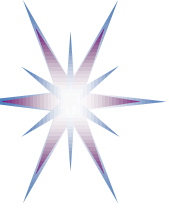
# SDI Architecture Layers

- **Layer D1:** Network infrastructure layer represented by the general purpose Internet infrastructure and dedicated network infrastructure
- **Layer D2:** Datacenters and computing resources/facilities
- **Layer D3:** Infrastructure virtualisation layer that is represented by the Cloud/Grid infrastructure services and middleware supporting specialised scientific platforms deployment and operation
- **Layer D4:** (Shared) Scientific platforms and instruments specific for different research areas
- **Layer D5:** Federation and Policy layer that includes federation infrastructure components, including policy and collaborative user groups support functionality
- **Layer D6:** Scientific applications and user portals/clients

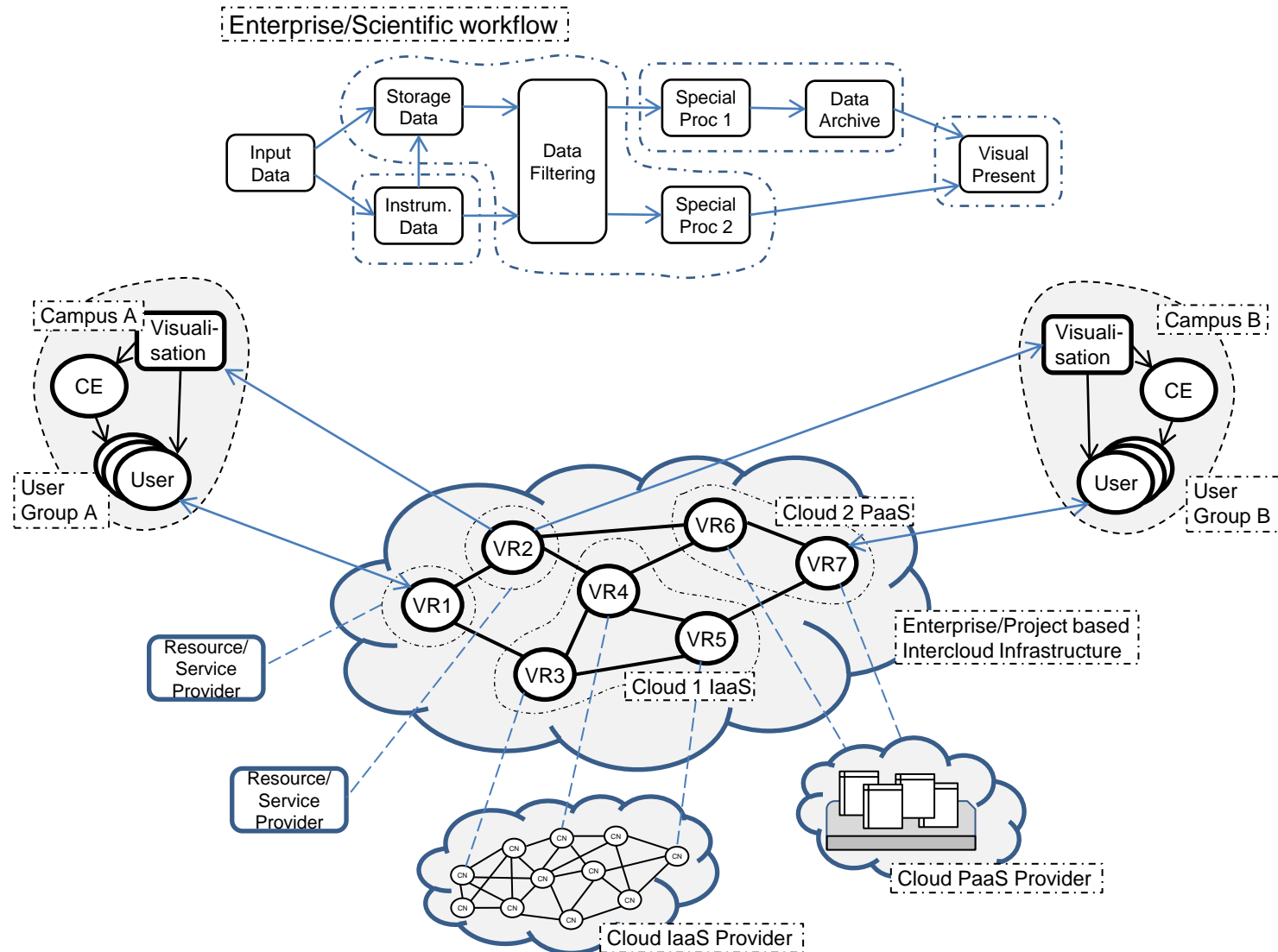


# SDI move to Clouds

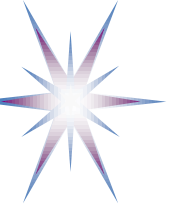
- Cloud technologies allow for infrastructure virtualisation and its profiling for specific data structures or to support specific scientific workflows
  - Clouds provide just right technology for infrastructure virtualisation to support data sets
  - *Complex distributed data require infrastructure*
- Cloud can provide infrastructure on-demand to scientific workflows
  - Similar to Grid but with benefits of the full infrastructure provisioning on-demand



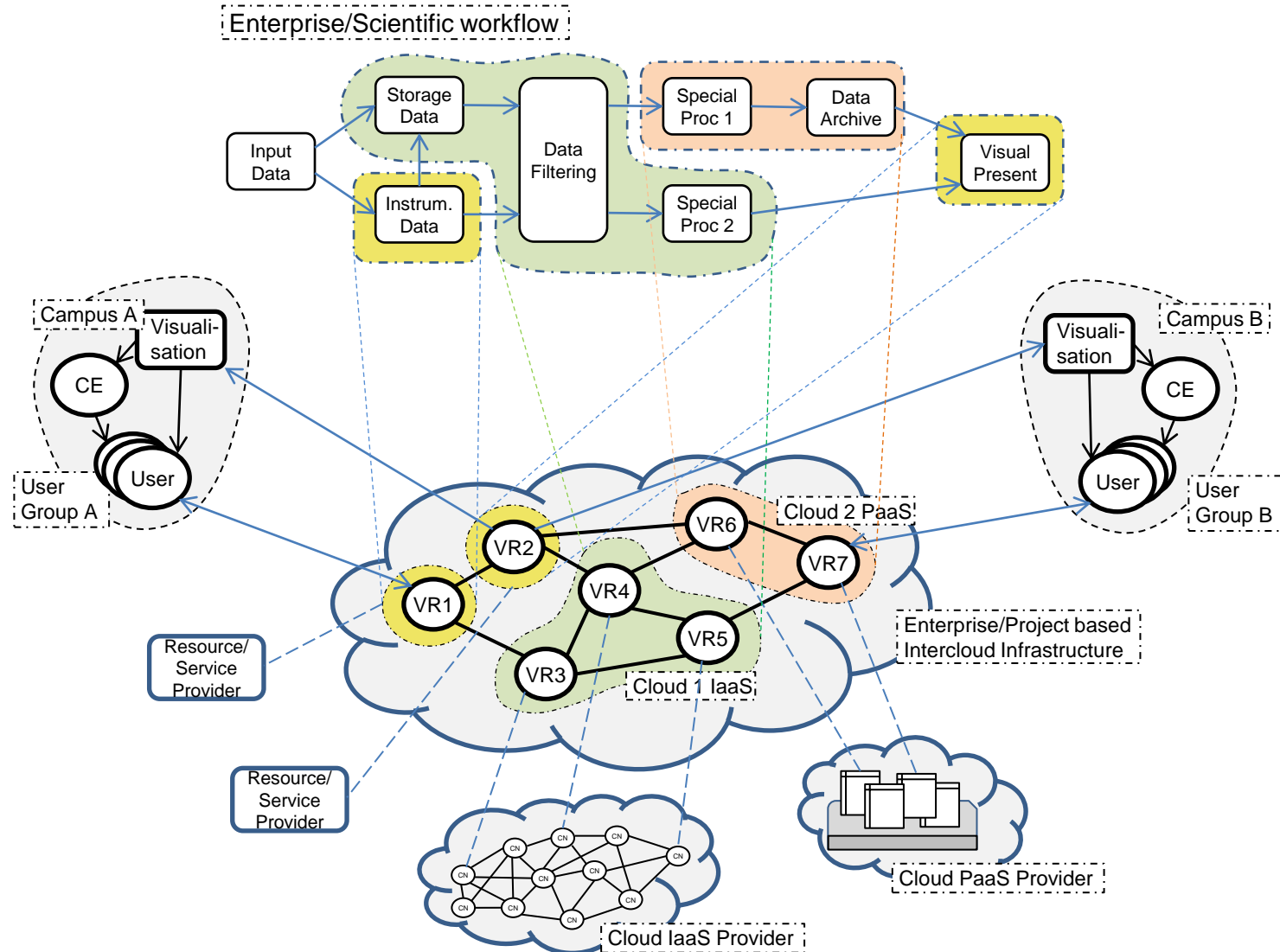
# General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure

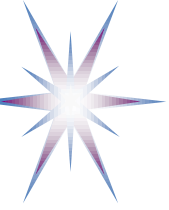




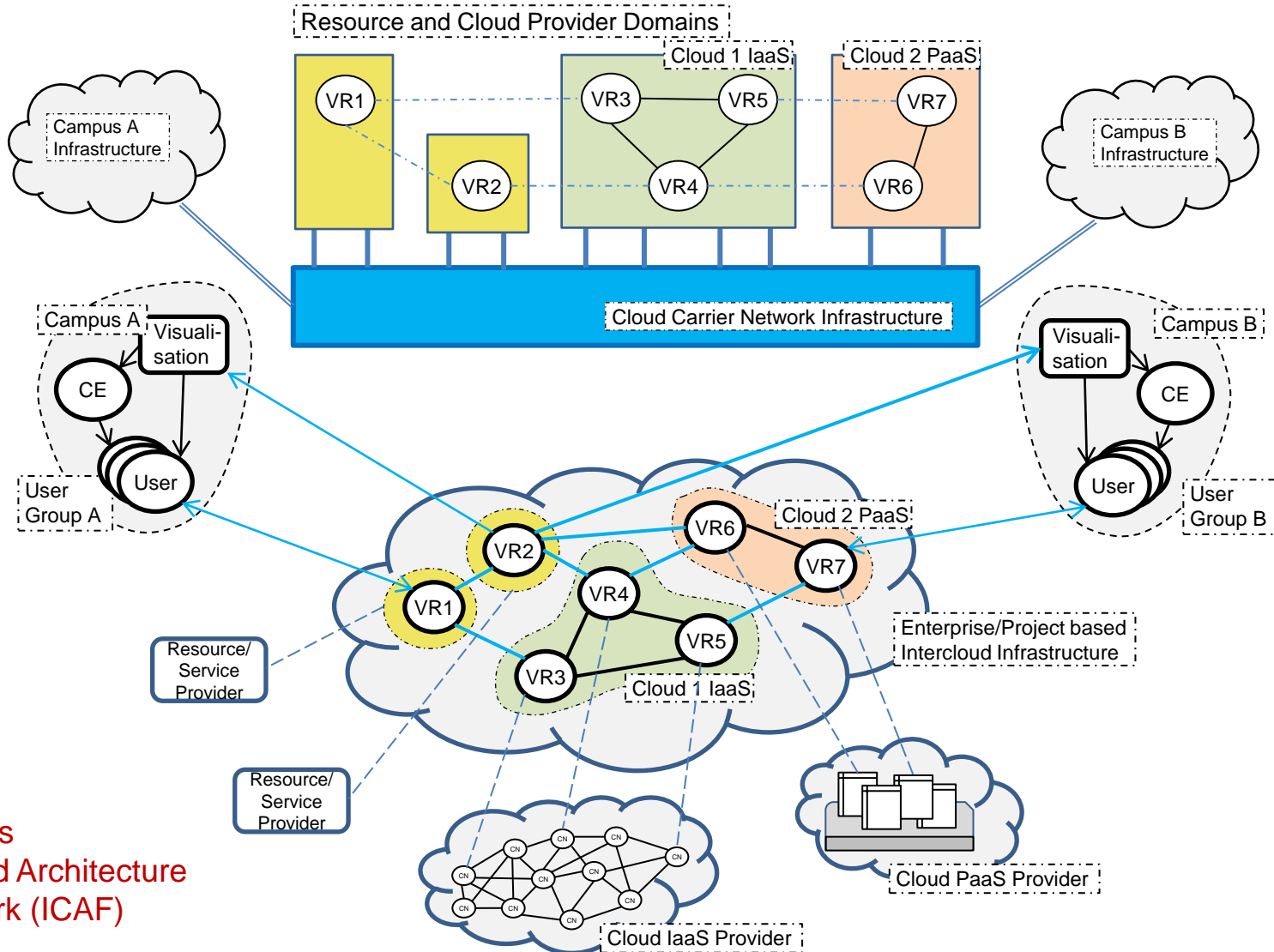


# General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure

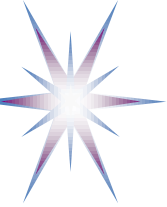




# General use case for infrastructure provisioning: Logical Infrastructure => Network Infrastructure (1)



Defined as  
InterCloud Architecture  
Framework (ICAF)

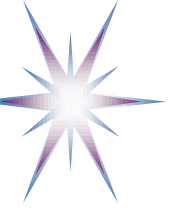


# InterCloud Architecture Framework (ICAF) Components

- **Multi-layer Cloud Services Model (CSM)**
  - Combines IaaS, PaaS, SaaS into multi-layer model with inter-layer interfaces
  - Including interfaces between cloud service layers and virtualisation platform
- **InterCloud Control and Management Plane (ICCMP)**
  - Allows signaling, monitoring, dynamic configuration and synchronisation of the distributed heterogeneous clouds
  - Including management interface from applications to network infrastructure and virtualisation platform
- **InterCloud Federation Framework (ICFF)**
  - Defines set of protocols and mechanisms to ensure heterogeneous clouds integration at service and business level
  - Addresses Identity Federation, federated network access, etc.
- **InterCloud Operations Framework (ICOF)**
  - RORA model: Resource, Ownership, Role, Action
  - Business processes support, cloud broker and federation operation

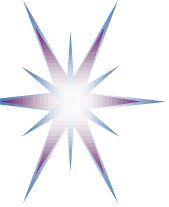
Intercloud Architecture for Interoperability and Integration, Release 1, Draft Version 0.5. SNE Technical Report 2012-03-02, 6 September 2012

<http://staff.science.uva.nl/~demch/worksinprogress/sne2012-techreport-12-05-intercloud-architecture-draft05.pdf>



# Need for new Scientific and Academic discipline

- New Scientific and Academic Discipline is needed for Big Data Science
  - *Included into declaration of the Rome EC Horizon2020 Consultation meeting (3-4 April 2012)*
- To address
  - ICT infrastructure building, operation and optimisation
  - Big data management and distributed computing
  - Metadata and semantics
  - Security and trustworthiness of Infrastructure and Data
- Need to educate/train new specialists in Big Data
  - Primarily for not ICT savvy community



# Questions and Discussion

---

Thank you for your attention.