Defining extra Vs: Value and Veracity

for Big Data 3V: Volume, Velocity, Variety

Big Data Technology Development 2012 Overview

Yuri Demchenko,
SNE Group, University of Amsterdam

3 January 2013, UvA, Amsterdam

Outline

- Research Data Alliance (RDA)
- Big Data definitions and variants
  - 3 V's of Big Data: Volume, Velocity, Variety
- Big Data technology drivers and challenges
- Adding more V's?
  - Peter Membrey's research on defining 4th V - Value
    - Use case: High Volume Low Value (HVLV) data for financial market feeds
  - Security and Trust: 5th V – Veracity
- Big Data Anatomy and Infrastructure issues
  - Big Data (multidimensional) models
  - Big Data Infrastructure components
  - Michael Stonebraker – database guru, SQL defender, NoSQL opponent
- Will Big Data definition and name change or evolve?
- Optional - Our recent papers and ongoing research

From where it is originated

- AAA Study: Study on AAA Platforms For Scientific data/information Resources in Europe - https://confluence.terena.org/display/aaastudy/AAA+Study+Home+Page
  - Final report published (de-personalised) http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf

European Commission

Advancing Technologies and Federating Communities

A Study
on Authentication
and Authorisation Platforms
For Scientific Resources
in Europe

FINAL REPORT
A study prepared for the European Commission
DG Communications Networks, Content & Technology

Digital Agenda for Europe

This study was carried out for the European Commission by

TERENA

UNIVERSITY OF AMSTERDAM

Internal identification
SMART-Nr 2011/0056

LEGAL NOTICE
By the Commission of the European Union, Communications Networks, Content & Technology Directorate-General,

Neither the European Commission nor any person acting on its behalf is responsible for the use which might be made of the information contained in the present publication.

The European Commission is not responsible for the external web sites referred to in the present publication.

The views expressed in this publication are those of the authors and do not necessarily reflect the official European Commission view on the subject.

© European Union, 2012

Reproduction is authorised provided the source is acknowledged.

Printed in Belgium

Big Data Science & Technology as the next technology focus

Scientific and Research Data – e-Science

- *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 (data supporting) infrastructure virtualisation

Big Data Challenges and Initiatives

- A Vision for Global Research Data Infrastructure (http://www.grdi2020.eu/)
  - Final Roadmap Report published
- Peta and Exa scale problems: Storage, Computing, Transfer/Network
  - International Exascale Software Project (http://www.exascale.org/)

- International Initiative "Research Data Alliance (RDA)" http://www.rd-alliance.org/ launched 2-3 Oct 2012, Washington
  - *To accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability*
  - Consolidated previous initiatives
    - Data Web Forum (DWF) initiated by NSF
    - DAITF – Data Access and Interoperability Task Force initiated by EUDAT project
  - Next meeting – RDA Official Launch, Gothenburg 18-20 March 2013
  - NIST Cloud and Big Data Workskhop 15-17 January 2013, Washington

Research Data Alliance – First Steps

- Organisation structure
  - International Steering Group (represent funding agencies in the US, EU and Australia)
  - RDA Council
  - Non-Governmental Steering Group (NGS)
    - Peter Wittenburg (Max Planck Institute for Psycholinguistics, Nijmegen)
    - Juan Bicarregui, Acting Director e-Science, STFC Rutherford Appleton Laboratory
    - Leif Laaksonen, Collaboration Director, CSC Finland
- Working Groups created
  - Harmonization and Use of PID Information Types
  - UPC (Universal Product Code) Code for Data
  - *Data Type Registries*
  - *Metadata*
  - Pub/Data Citation/Linking
  - Data Foundation and Terminology
  - Practical Policy
  - Legal Interoperability
  - Defining Urban Data Exchange for Science
  - Marine Data Harmonization
  - *Repository Audit and Certification*
  - The Engagement Group

Missed topics:

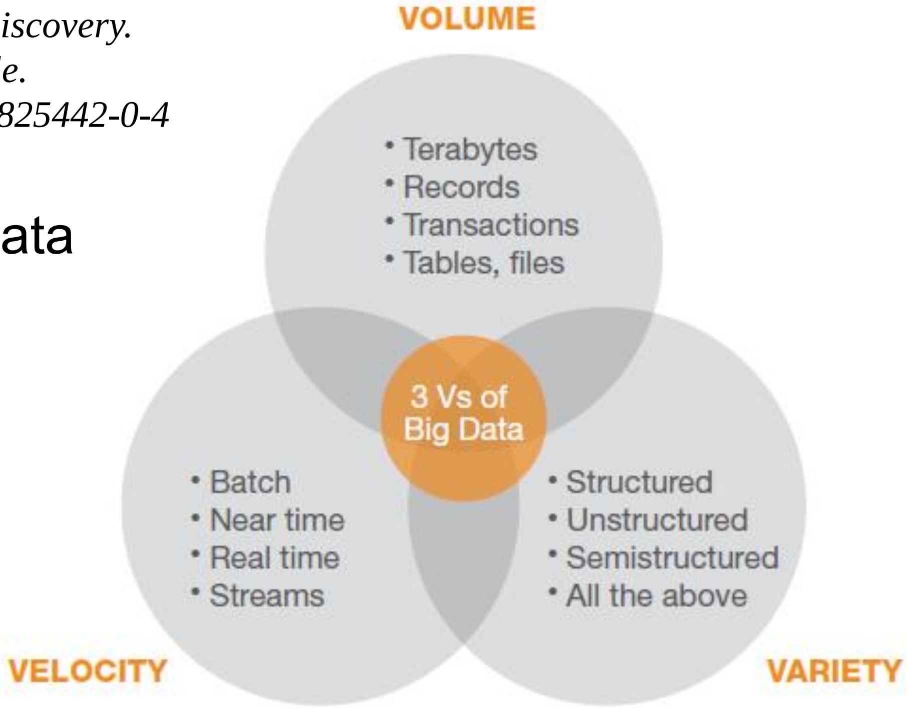
- Infrastructure
- Security and Identity Management
- Use cases and BCP

Big Data Definition

- Termed as the Fourth Paradigm *)
  *"The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration." (Jim Gray, computer scientist *)*

**) The Fourth Paradigm: Data-Intensive Scientific Discovery.*
Edited by Tony Hey, Stewart Tansley, and Kristin Tolle.
Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4

- Commonly accepted 3 V's of Big Data
  - Volume
  - Velocity
  - Variety



Claiming BD 3V Origin by Doug Laney

- Doug Laney – now with Gartner: VP Research, Business Analytics and Performance Management
  - Article “Deja VVVu: Others Claiming Gartner’s Construct for Big Data”
    - http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/
  - *3-D Data Management: Controlling Data Volume, Velocity and Variety (6 February 2001)*
-

Volume, Velocity, Variety - Examples

- Volume – Terabyte records, transactions, tables, files.
  - A Boeing Jet engine spews out 10TB of operational data for every 30 minutes they run
  - Hence a 4-engine Jumbo jet can create 640TB on one Atlantic crossing. Multiply that to 25,000 flights flown each day and we get the picture
- Velocity – batch, near-time, real-time, streams.
  - Today’s on-line ad serving requires 40ms to respond with a decision.
  - Financial services (i.e., stock quotes feed) need near 1MS to calculate customer scoring probabilities.
  - Stream data, such as movies, need to travel at high speed for proper rendering.
- Variety – structures, unstructured, semi-structured, and all the above in a mix.
  - WalMart processes 1M customer transactions per hour and feeds information to a database estimated at 2.5PB (petabytes).
  - There are old and new data sources like RFID, sensors, mobile payments, in-vehicle tracking, etc.

Big Data definition – Variations and extensions

- Ed Dumbill, program chair for the O’Reilly Strata Conference
  - Big Data - “data that exceeds the processing capacity of conventional database systems. *The data is too big, moves too fast, or doesn’t fit the structures of your database architectures*. To gain value from this data, you must choose an alternative way to process it.”
- IBM: to add another V - Veracity (confirming to truth/fact)
  - 1/3 business leaders make decisions based on information that don’t fully trust. How can you act upon information if you don’t trust it?
  - Establishing trust in big data presents a huge challenge as the variety and number of sources grows
  - *Ons feilbare denken - Thinking, fast and slow, by Daniel Kahneman*
- 3 I’s of Big Data from Forbes Research (Dave Feinleib referring to Vance Loiselle, Sumo Logic)
  - Immediate – in the sense that you need to do something about it now
  - Intimidating – what if you don’t?
  - Ill-defined – what is it, anyway?

Big Data technology drivers (1)

- Modern e-Science in search for new knowledge
  - Scientific experiments and tools are becoming bigger and heavily based on data processing and mining
- Traditional data intensive industry
  - Genomic research, drugs development, Healthcare
  - High-tech industry, CAD/CAM, weather/climate, etc.
- Consumer facing companies like Google and Facebook have driven many of the recent advances in Big Data efficiency
  - Facebook has some 900+ million users and is still growing
  - Google handles number of search queries at 3 billion per day
  - Twitter handles some 400 million tweets per day count for 12 terabytes per day
    - Used also for market sentiments prediction
  - Power companies: process up to 350 billion annual meter readings to better predict power consumption

Big Data technology drivers (2)

- Technology loop (known as Jevons Paradox)
  - Increased efficiency to process current demand will create new uses and increase demand even more
- Processes/activity data recording and analysis
  - Flight data, log data, intelligence, traffic
- Business (retail) uses Big Data technologies “to search” for customers
  - Modern business concept (multi-channel) of delivering directly to customers requires prediction of customer behavior
    - Data volumes – What cause(s) and what effect?
  - Big Data gives organisations a fighting chance in the battle over the customer

Big Data technology drivers (3) - Advertising

- "… this new course of big data, gleaned from a wealth of unstructured information on the web, has the ability to turn advertising on its head— at least enough to make media people rethink algorithms for maximizing performance." *HessieJ.com*
  - *Traditional Ad Model: User profiles*
  - *More Sophisticated Ad Model: Behavioral targeting - "smart ads"*
  - *Future Ad Model: Enter Social Data*
- Case in point:
  - Mary Brown searches for information about a future trip to Hawaii
    - She also goes to travel sites, reads hotel reviews and has excitedly spoken to close friends on Twitter and Facebook about her plans and preparations
  - Now we have not only recent behavioral activity *where she's been on the interne*t, but we also are aware of her conversations that validate her behavior
  - It is safe to assume that Mary will "definitely" be going to Hawaii
  - What this information does for a travel company?
    - They now have ***MORE*** information on that user that will allow them to not only *serve* an ad, or even *respond* to that user with relevant offers, but ***DO*** so with a certain degree of confidence that Mary will at the very least click on the ad.

Big Data technology drivers (3a) – Service Delivery

- Consumer products and services delivery
  - Netflix already captures movie genre preferences by the user and makes recommendations based on recent shows/movies watched
    - Announced $2mln prize for effective customer targeting in 2003
  - It is already capturing which devices the user is watching recent programs/shows and when
    - Marrying that data with GetGlue (news feed on movies), for example, validates the original information and supplements the user information with commenting, share data as well as potential prospects
  - When combined and correlated, these snippets provide insight that now allows Netflix to optimize the movie offering to you to keep you a satisfied customer
  - It can also capture the comments and shares from those watching the movie in order to drive messaging to attract new users

Big Data technology drivers (3b) – Managing public campaigns, e.g. election, IPR

- The rise of public opinion stored in platforms like Twitter, Google, Facebook, etc. provide enough intelligence to influence the campaign development, timing, geography and even the colour of the campaign signs
  - Twitter was a major source of data aggregation for the Republican Race in the US

Big Data technology drivers (4) - Emerging

- Social media itself – share and socialise/collaborate
- Workplace improvement
  - Means more data will be collected and monitored on the personnel
- Healthcare, health/physiological and medical information
  - Human health monitoring – not just for ill or aged people

Big Data Challenges - Technological

- How to scale up and down (scale or shrink)?
  - Primarily database issues
  - SQL scales easy up but not easy scales down if demand decreases
    - NoSQL (Not only SQL) can partly address this issue
    - SQL has complex syntax, strong schema typing, performance
  - NoSQL is more flexible to adopt to new biz processes
    - Primarily but not just key-value or document-based
- Data structures and data models
  - To respond to specific use cases and operations over data
- Data mining/data intelligence algorithms
  - To handle/discover new data structures and multi-type data relations
  - Human/behavioral/social targeted data analysis (means fuzzy/biased)
- Infrastructure support for storing, moving data, on-demand processing
  - Is Cloud Computing a right technology? Any alternative?
  - Highspeed network infrastructure, on-demand provisioning
- Big Data security, trustworthiness and data centric security

Big Data Challenges – Socio-technological

- Extending big data outreach/perimeter
  - Technology will boom if there is sufficient customer and user base
    - Currently majority of Big Data consumers are big companies
      - Although we are contributing with feeding our activity/usage log data
  - Move big data from big companies to user and homes
    - Smart homes, sensors and devices
      - Without sensors and devices human can not create or use big data
      - Smart visualisation can solve a problem of using/acting on big data
- Lowering entry level to use Big Data
  - You should not be a data expert to use Big Data
  - Needs for scalable configurable tools

Big Data: Existing technologies (1)

- Big Data storage and data base solutions
  - – Google File System, BigTable
  - – New types of databases?
- Big Data computing
  - – MapReduce, Google’s Dremel
  - – MPI and Computer Grids
  - – To be capable to run in a distributed environment closer to data stores and sources
- Network for Big Data
  - – Capacity – Yes
  - – New protocols? – Routing, streaming, load balancing, etc.

Big Data: Existing technologies (2)

- Social networks and human driven/centric collaboration and sharing environment
  - Facebook is a heuristic implementation to human’s crave to share, socialise and “expand”
    - Not perfect, to evolve or be overtaken by a new more open technology to support human’s socialising activity
    - Current new developments at Facebook
      - From Like to Open Graph as a behavioral technology
      - Payment system
- Smartphones, tablets and BYOD (Bring Your Own Device) – to be supported by employer IT
  - Hub for human/body sensors and wearable devices

Database Issues

- Data structures and file systems
  - Block storage, file storage, network storage
  - E.g., Lustre and LVM, Google File System
- Hadoop/MapReduce and NoSQL
- SQL vs NoSQL – Relational vs non-relational
  - Google Bigtable architecture
    - HBase – Hadoop oriented
  - Amazon’s Dynamo architecture
  - Cassandra database developed by Facebook (Dynamo based)
    - Cassandra is able to store 2 million columns in a single row
- MongoDB – document oriented (in memory DB)
  - Data are structured as JSON-like constructs (BSON) with dynamic schemas
  - The most popular NoSQL DBMS
  - Easily combined with MapReduce
  - Powers In-Memory-Computing – a paradigm for Big Data real-time processing (actively developed by SAP)

Michael Stonebraker (MIT since 2001)

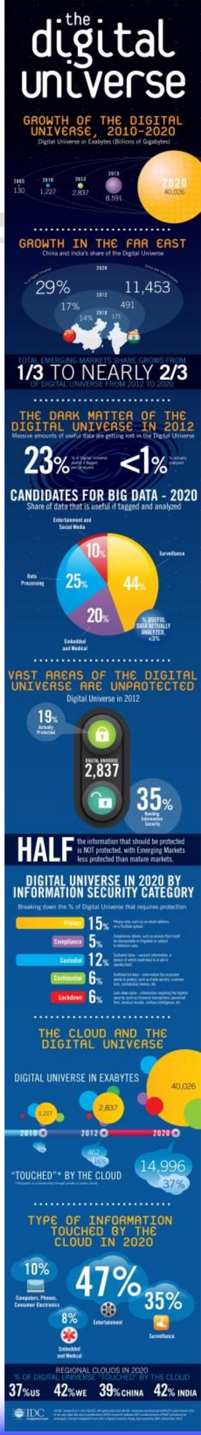
- Stonebraker is critical about NoSQL since 2003
- Aurora and StreamBase (2003)
  - Data management for streaming data, using a new data model and query language. In Aurora, data is "pushed", arriving asynchronously from external data sources (such as stock ticks, news feeds, or sensors.) The output is itself a stream of results (such as windowed averages).
- C-Store and Vertica (2003)
  - Parallel, shared-nothing column-oriented DBMS for data warehousing. By dividing and storing data in columns, C-Store is able to perform less I/O and get better compression ratios than conventional database systems that store data in rows.
- Morpheus and Goby (2006)
  - Data integration system which relies on a collection of "transforms" to mediate between data sources. Morpheus makes it possible to search for and compose multiple transforms to provide a new service or a unified view of several services.
- H-Store and VoltDB (2009)
  - Distributed main-memory OLTP system designed to provide very high throughput on transaction processing workloads.
- SciDB
  - Open-source DBMS specially designed for scientific research applications

Foreseen Big Data Innovations in 2013

- Real-Time Hadoop
  - Google’s Dremel-like solutions that will allow real-time queries on Big Data and be open source
- Cloud-Based Big Data Solutions
  - Amazon’s Elastic Map Reduce (EMR) is a market leader
  - Expected new innovative Big Data and Cloud solutions
- Big Data Appliances (also for home)
  - Raspberry Pi and home-made GPU clusters
  - Hardware vendors (Dell, HP, etc.) pack mobile ARM processors into server boxes
  - Adepteva's Parallella will put a 16-core supercomputer into for $99
- Distributed Machine Learning
  - Mahout iterative scalable distributed backpropagation machine learning and data mining algorithm
  - New algorithms Jubatus, HogWild
- Easier Big Data Tools
  - Open Source and easy to use drag-and-drop tools for Big Data Analytics to facilitate the BD adoption
  - Commercial examples: Radoop = RapidMiner + Mahout, Tableau, Datameer, etc.

Predictions for 2013 and as far as 2020

- Main technologies and areas
  - Smartphones and tablets
  - E-Medicine and Healthcare
  - Big Data
  - Cloud: shift from private to public
- Big Data and Cloud Infographic http://www.cloudtweaks.com/2013/01/cloud-infographic-big-data-universe/



Will Big Data term sustain? – Other names

- Big Analytics, Big Data Analytics
- Data Analytics, Intelligent Analitics
  - Missed infrastructure component
- New concepts related to Big Data
  - Disposable Data – in contrary to data supposed to be stored

Possible Research topics and Papers

- 4th V – Value -> with Peter Membrey
- 5th V – Veracity -> Trustworthiness and security
- Big Security for Big Data
- Defining the nature of the Big Data
- Scientific Data Lifecycle Management Model
- Federated AAI for Scientific Data Infrastructure (SDI)
- Cloud based Big Data infrastructure
- Data Management and Security in Clouds

Questions and Discussion

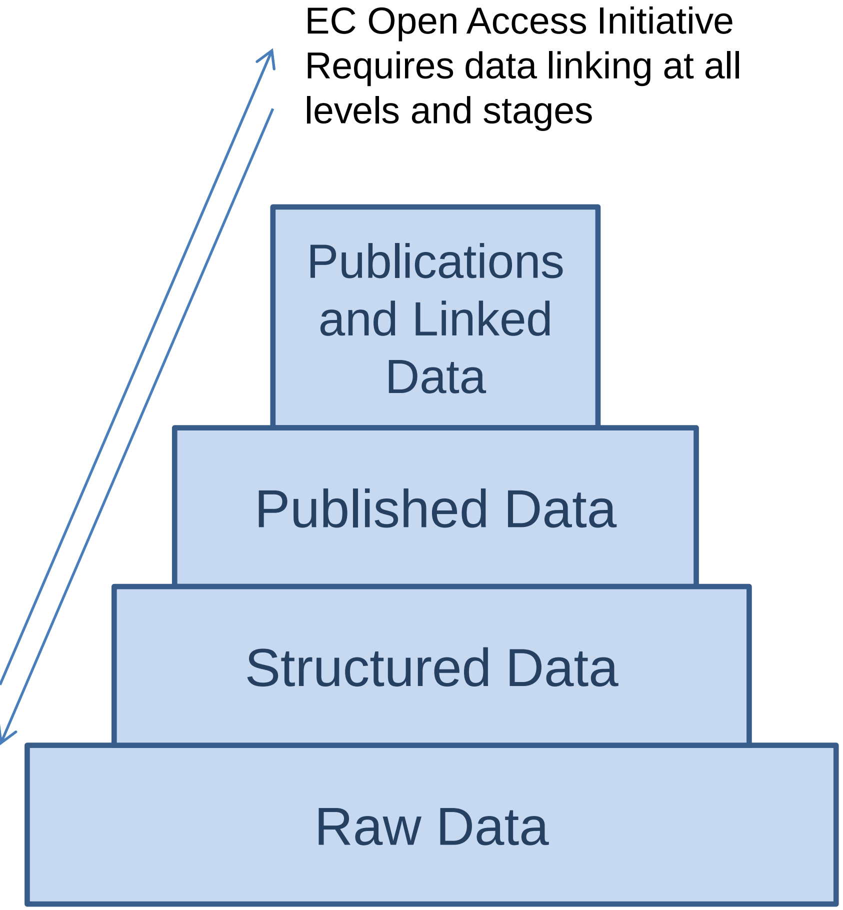
Recent research at SNE/UvA on defining Big Data Infrastructure - 2012

- AAA technologies for Scientific Data Infrastructure (SDI)
- Big Data challenges for Scientific Data Infrastructure
- Scientific Data Lifecycle Management (SDLM) Model
- Cloud based infrastructure services provisioning for Scientific Data projects and collaboration

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 i***nfrastructures 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



EC Open Access Initiative Requires data linking at all levels and stages

Publications and Linked Data

Published Data

Structured Data

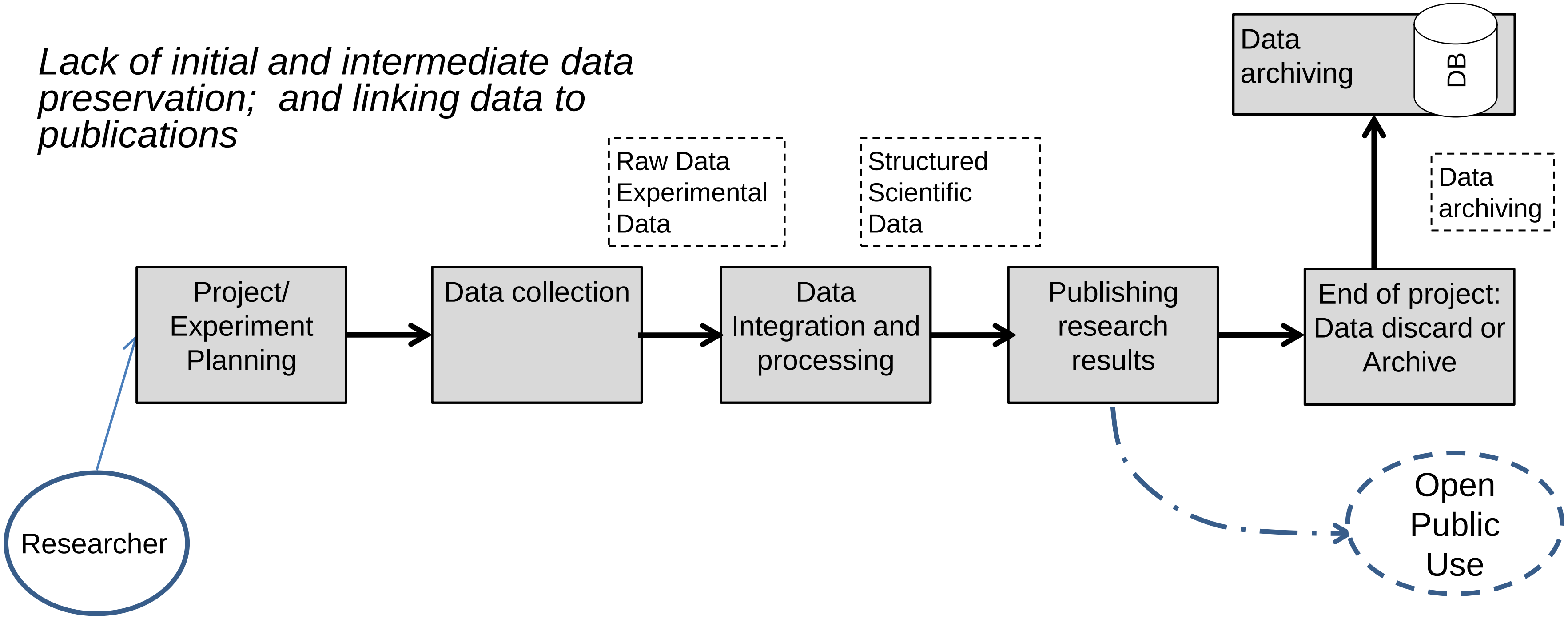
Raw Data

- **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

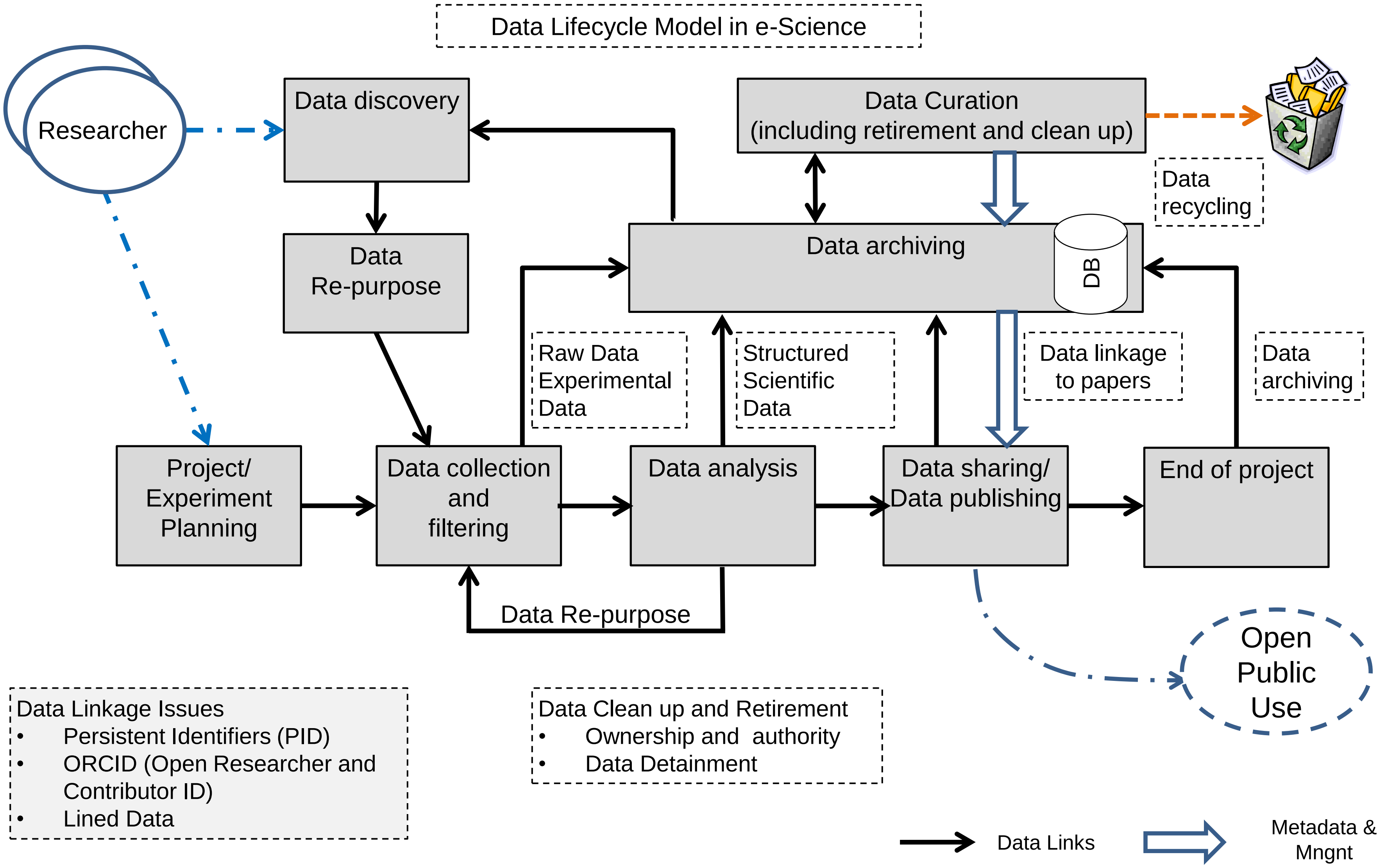
- Data collection
- Data processing
- Publishing research results
- Discussion
- Data and publications archiving

Lack of initial and intermediate data preservation; and linking data to publications



Data archiving
DB
Raw Data Experimental Data
Structured Scientific Data
Data archiving
Project/ Experiment Planning
Data collection
Data Integration and processing
Publishing research results
End of project: Data discard or Archive
Open Public Use
Researcher

Data Lifecycle Model in e-Science



Data Lifecycle Model in e-Science

Researcher

Data discovery

Data Re-purpose

Data Curation (including retirement and clean up)

Data recycling

Data archiving

DB

Raw Data Experimental Data

Structured Scientific Data

Data linkage to papers

Data archiving

Project/ Experiment Planning

Data collection and filtering

Data analysis

Data sharing/ Data publishing

End of project

Data Re-purpose

Open Public Use

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

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

Data Links

Metadata & Mngnt

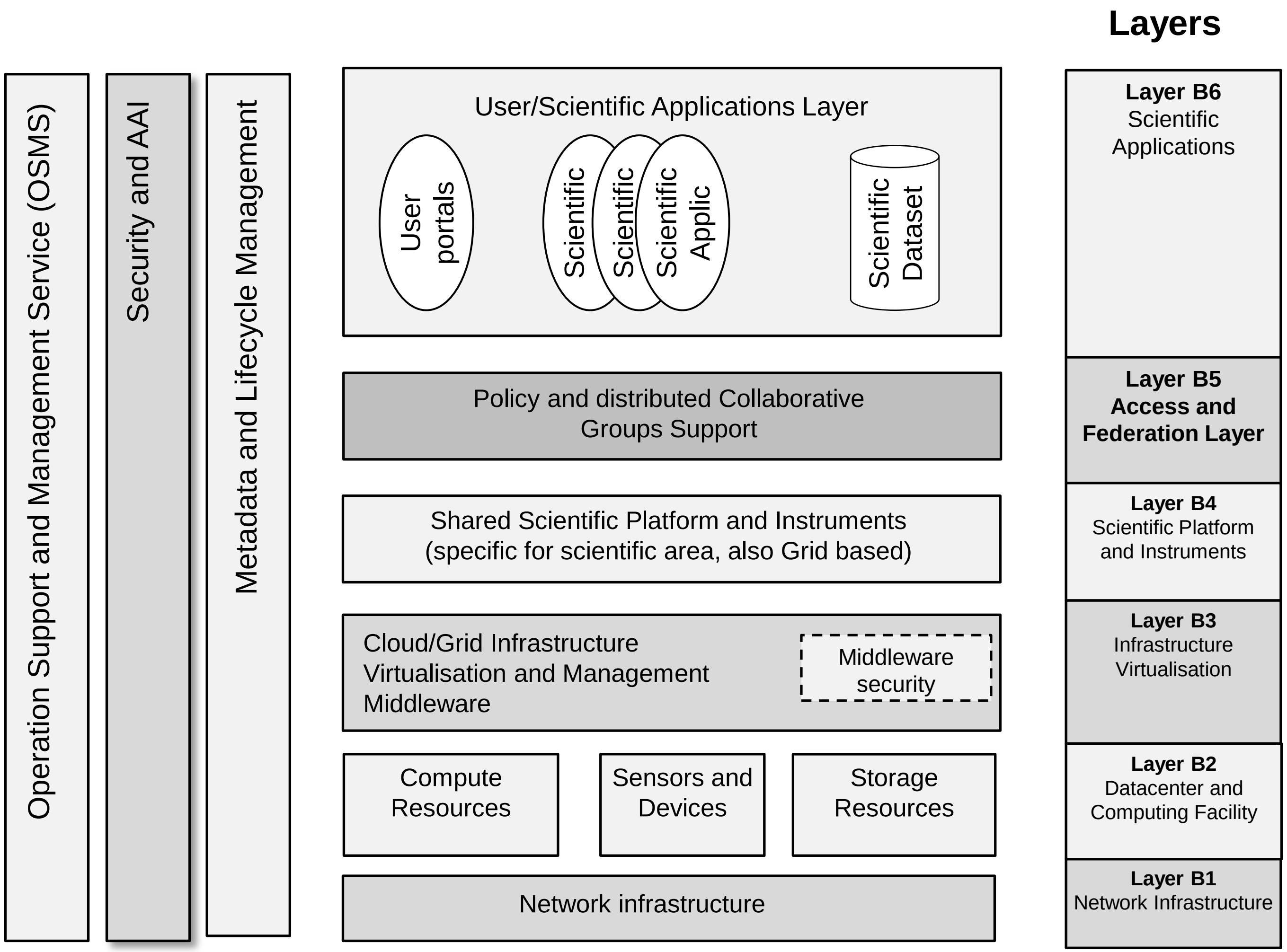
General requirements to SDI for emerging Big Data Science

- 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 Security

- 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
  - New actors in SDI (and Big Data Infrastructure)
    - Subject of data (e.g. patient, or scientific object/paper)
    - Data Manager (doctor, seller)
- Security and Access Control and Accounting Infrastructure (ACAI)
  - Trust management infrastructure
  - Authentication, Authorisation, Accounting
    - Supported by logging service
  - Extended to support data access control and operations on data

SDI Architecture Model



Operation Support and Management Service (OSMS)

Security and AAI

Metadata and Lifecycle Management

User/Scientific Applications Layer

User portals

Scientific

Scientific

Scientific Applic

Scientific Dataset

Policy and distributed Collaborative Groups Support

Shared Scientific Platform and Instruments (specific for scientific area, also Grid based)

Cloud/Grid Infrastructure Virtualisation and Management Middleware

Middleware security

Compute Resources

Sensors and Devices

Storage Resources

Network infrastructure

Layers

Layer B6 Scientific Applications

Layer B5 **Access and Federation Layer**

Layer B4 Scientific Platform and Instruments

Layer B3 Infrastructure Virtualisation

Layer B2 Datacenter and Computing Facility

Layer B1 Network Infrastructure

Technologies and solutions

Scientific specialist applications
Library resources

Network Infrastructure
eduGAIN, REFEDS, VOMS, InCommon

PRACE/DEISA

Grid/Cloud

Clouds

Autobahn, eduroam

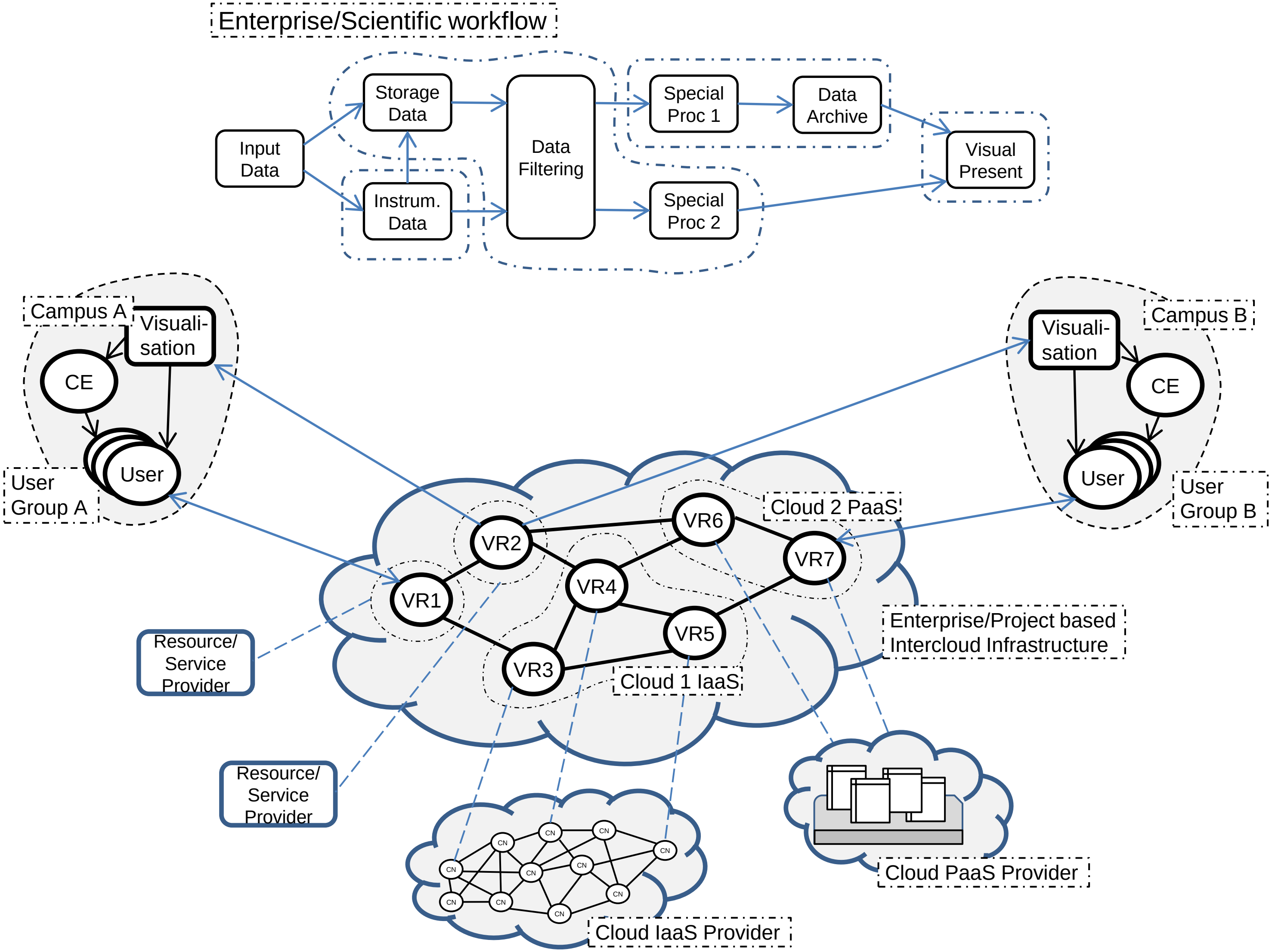
SDI Architecture Layers

- **Layer D1**: Network infrastructure layer represented by the general purpose Internet infrastructure and dedicated network infrastructure
- **Layer D2**: Datacenters and computing resources/facilities, including sensor network
- **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**: Access Infrastructure Layer: Federation infrastructure components, including policy and collaborative user groups support functionality
- **Layer D6**: Scientific applications and user portals/clients

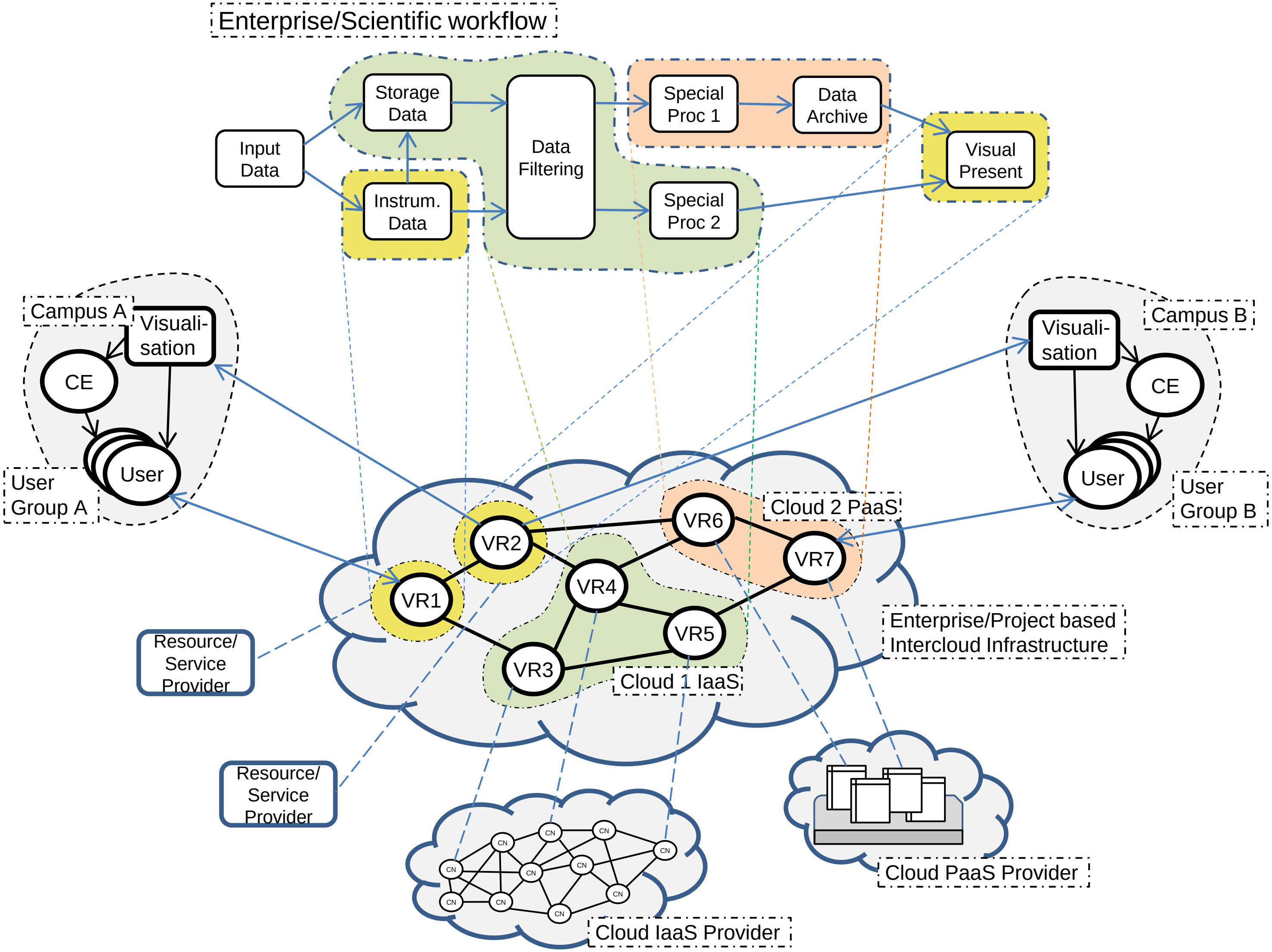
SDI move to Clouds

- 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*
    - *Demand for inter-cloud infrastructure*
- Cloud can provide infrastructure on-demand to support project related scientific workflows
  - Similar to Grid but with benefits of the full infrastructure provisioning on-demand

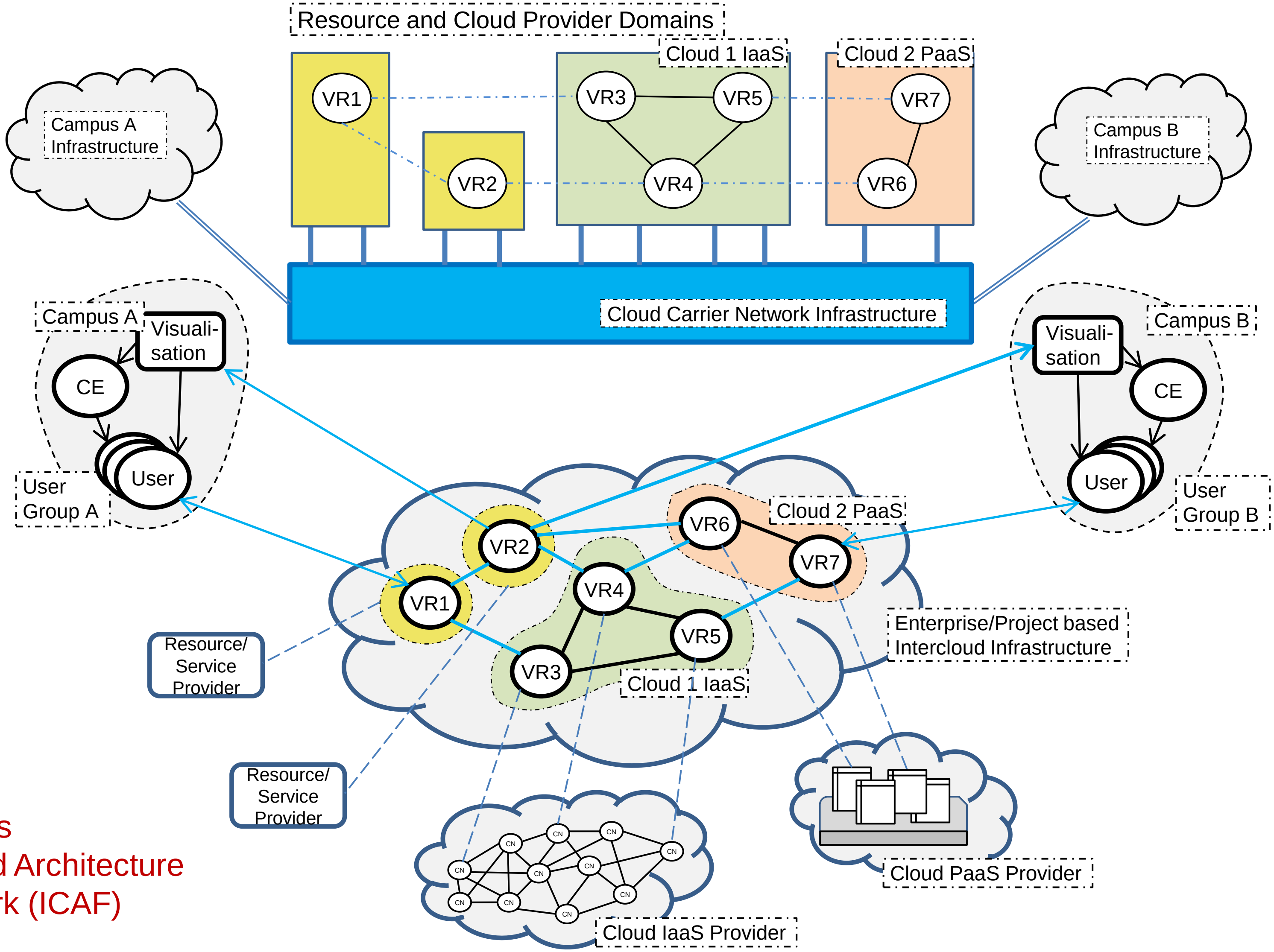
General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure



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)

Big Data and Intercloud Research topics

- Mapping from scientific workflow to inter-cloud
- Data structures and supporting infrastructure
- Cloud infrastructure support for Big Data security and trustworthiness (for generically distributed scenarios)
  - Authenticity, authorisation, delegation
  - Trust, validity
  - Accounting, auditing
  - Privacy