Addressing Big Data Issues in Scientific Data Infrastructure

Yuri Demchenko,
SNE Group, University of Amsterdam

21 May 2013, BDDAC2013 Workshop
CTS2013 San Diego

Outline

- Background to Big Data research at SNE/UvA
- Big Data definition
  - 5 V's of Big Data: Volume, Velocity, Variety, Value, Veracity
    - Use case: High Volume Low Value (HVLV) data for financial market feeds
  - Will Big Data term and definition change/evolve?
- Big Data technologies landscape
- Big Data and Data Intensive e-Science
  - Scientific Data Lifecycle Management
- Scientific Data Infrastructure (SDI) for Big Data
  - Cloud and Intercloud based platform for SDI
- Infrastructure research on SDI for Big Data
- Discussion

Background to Research on Big Data

- System and Network Engineering Group at the University of Amsterdam
  - Optical and high performance networks
  - E-Science and Collaborative applications
  - Computer Grids and Cloud Computing
  - Security, access control, trust management
- Future Scientific Data Infrastructure
  - Fusion of industry Big Data and Data Intensive Science
- Big Data requires infrastructure
  - High performance computing and network
  - Access infrastructure and collaborative environment
  - Clouds (and Grids) is not a full answer

Big Data and Data Intensive Science & Technology - The next technology focus

Scientific and Research Data – e-Science

- *Big Data is/has becoming the next buzz word*
  - Not much academic research and papers *– Dive into blogs and tweets*
- 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
  - First meeting – RDA Official Launch, Gothenburg 18-20 March 2013
  - Next meeting 15-17 September 2013, Washington

Big Data Definition (1)

- IDC definition (conservative and strict approach) of Big Data:
  "A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis"

- Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.
  - From "The Big Data Long Tail" blog post by Jason Bloomberg (Jan 17, 2013). http://www.devx.com/blog/the-big-data-long-tail.html

- "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."
  - Ed Dumbill, program chair for the O'Reilly Strata Conference

- 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)

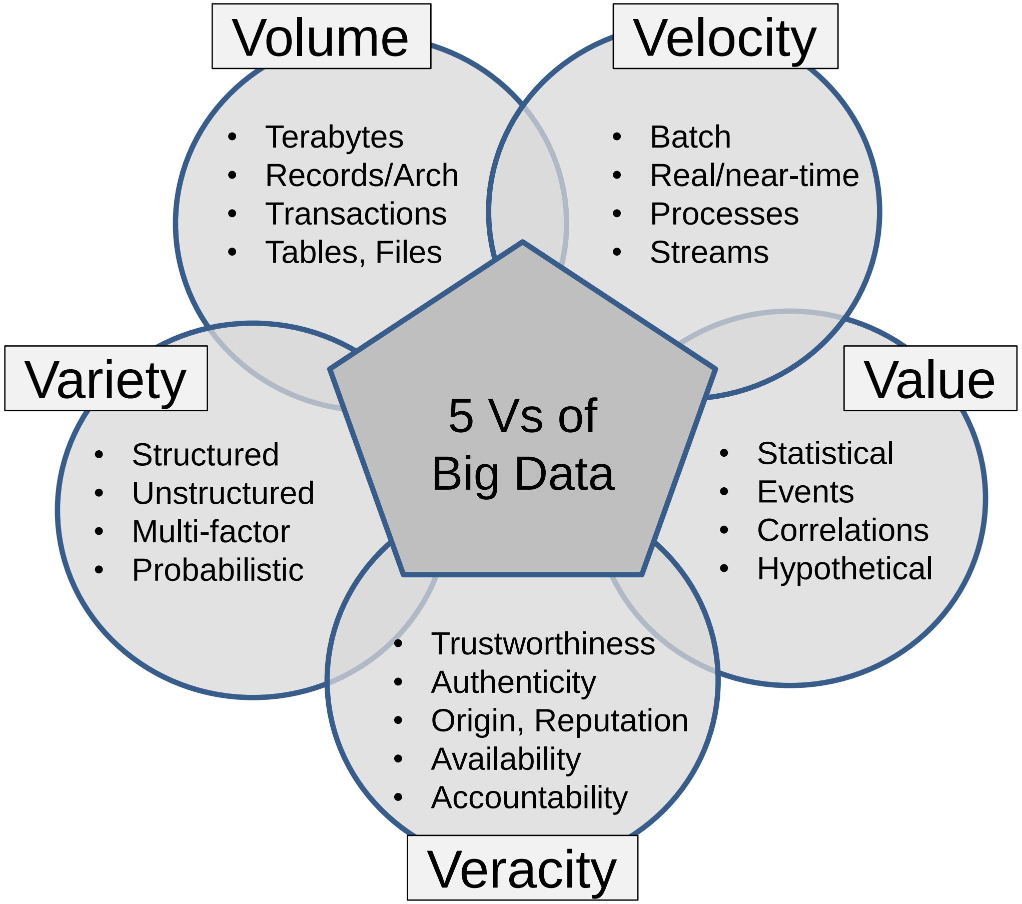
*) *The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft, 2009.*

Big Data Definition (2) - Motivation

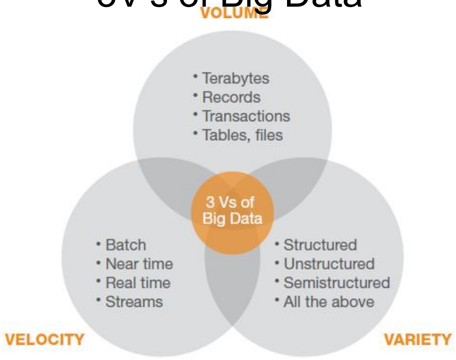
- Why Big Data based decision making support is so important?
  - Read book *"Thinking. Fast and Slow"* (2011) by Nobel Memorial Prize winner in Economics Daniel Kahneman
    - Our thinking is systematically mistaken when it comes to (subconscious) statistical data assessment
    - Dutch translation *"Ons feilbare denken"* ("Our mistaken thinking")

- Blog article by Mike Gualtieri from Forrester:
  "Firms increasingly realize that [big data] must use predictive and descriptive analytics to find nonobvious information to discover value in the data. Advanced analytics uses *advanced statistical, data mining and machine learning algorithms* to dig deeper to find patterns that you can't see using traditional BI (*Business Intelligence*) tools, simple queries, or rules."

5 V's of Big Data



Commonly accepted 3V's of Big Data



Volume, Velocity, Variety – Examples e-Science

- Volume – Terabyte records, transactions, tables, files.
  - LHC – 5 PB a month (now is under re-construction)
  - LOHAR – 5 PB every hour, requires processing asap to discard non-informative data
    - Other astronomy research
  - Genome research
  - Earth, climate and weather data
- Velocity – batch, near-time, real-time, streams.
  - LNC ATLAS detector collect about xTB data from X sensors during the collision time about 1? ms
  - What other research require highspeed data
- Variety – structures, unstructured, semi-structured, and all the above in a mix
  - Biological and medical, facial research
  - Human, psychology and behavior research
  - History, archeology and artifacts

Volume, Velocity, Variety – Examples Industry

- Volume – Terabyte records, transactions, tables, files.
  - A Boeing Jet engine produce 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 – Veracity

- Trustworthiness and Reputation -> Integrity
- Origin, Authenticity and Identification
  - Identification both Data and Source
  - Source: system/domain and author
  - Data linkage (for complex hierarchical data, data provenance)
- Availability
  - Timeliness
  - Mobility (mobile/remote access; from other domain – roaming; federation)
- Accountability
  - A kind of pro-active measure to ensure data veracity

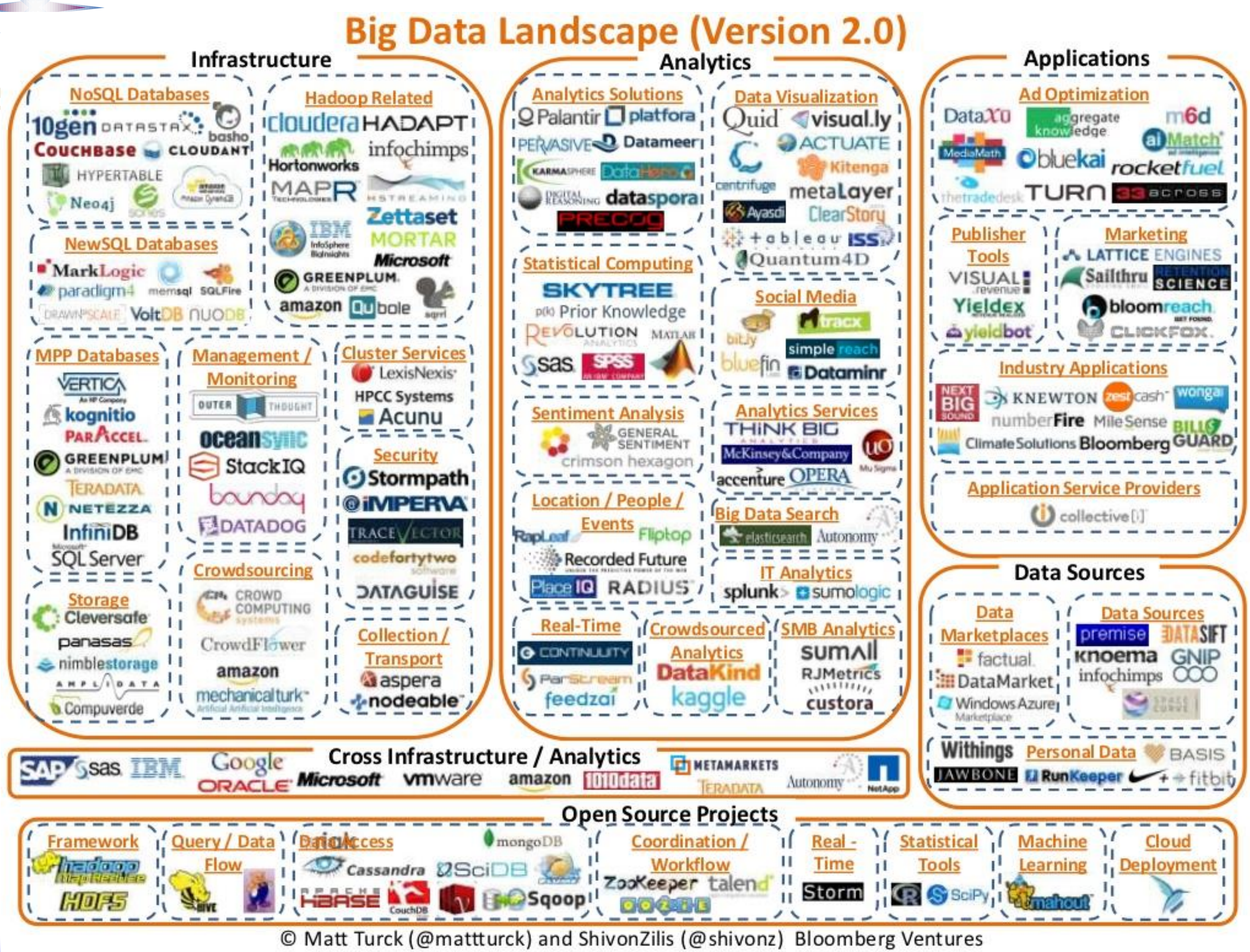
Big Data Challenges - Technological

- 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?
  - High speed 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 and Privacy issues
  - Digital footprint and re-identification

http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/



Big Data Infrastructure Components

- Cloud base infrastructure services for data centric applications (storage, compute, infrastructure/VM management)
- Cluster services
- Hadoop related services and tools
- Specialised data analytics tools (logs, events, data mining, etc.)
- Databases/Servers SQL, NoSQL
- MPP (Massively Parallel Processing) databases
- Big Data Management
- Registries, Indexing/search, semantics, namespaces
- Security infrastructure (access control, policy enforcement, confidentiality, trust, availability, privacy)
- Collaborative environment (groups management)

Will Big Data term sustain? – Other names

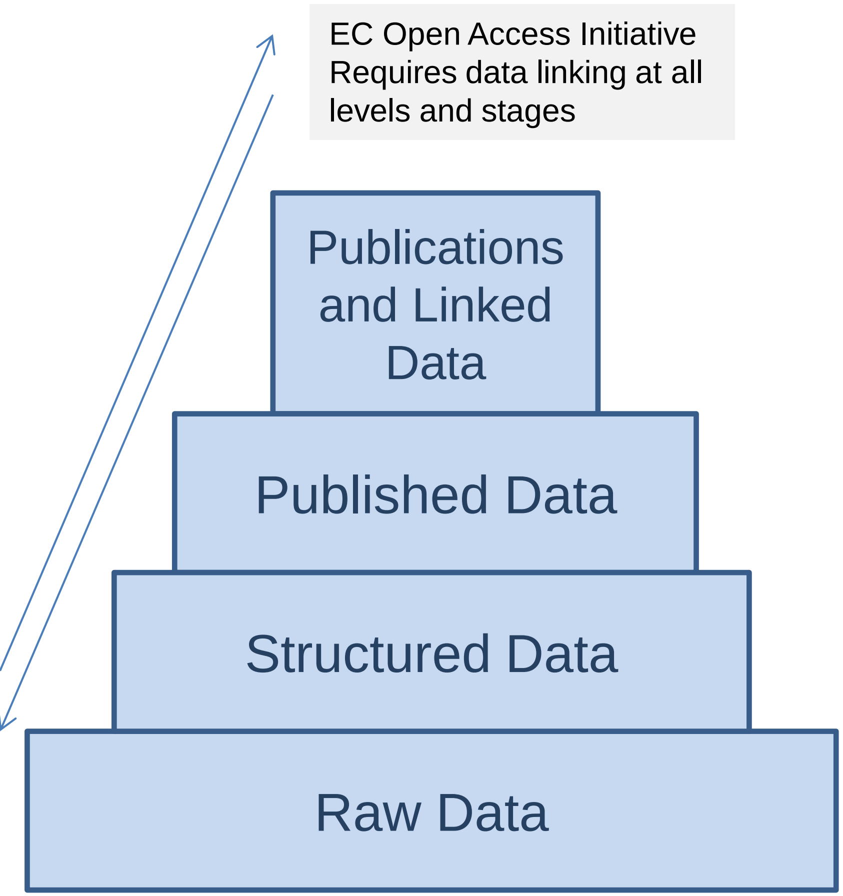
- Big Analytics, Big Data Analytics
  - To avoid the term itself becomes a “fetish”
- Data Analytics, Intelligent Analytics
  - Missed infrastructure component
- Big Data vs Data Intensive Science
  - e-Science is based on and involves wide cooperation between researchers
- New concepts related to Big Data
  - Disposable Data – in contrary to data supposed to be stored
  - Non-deterministic nature of the scientific study objects
    - From natural science to economics and social science

Big Data and Data Intensive Science

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

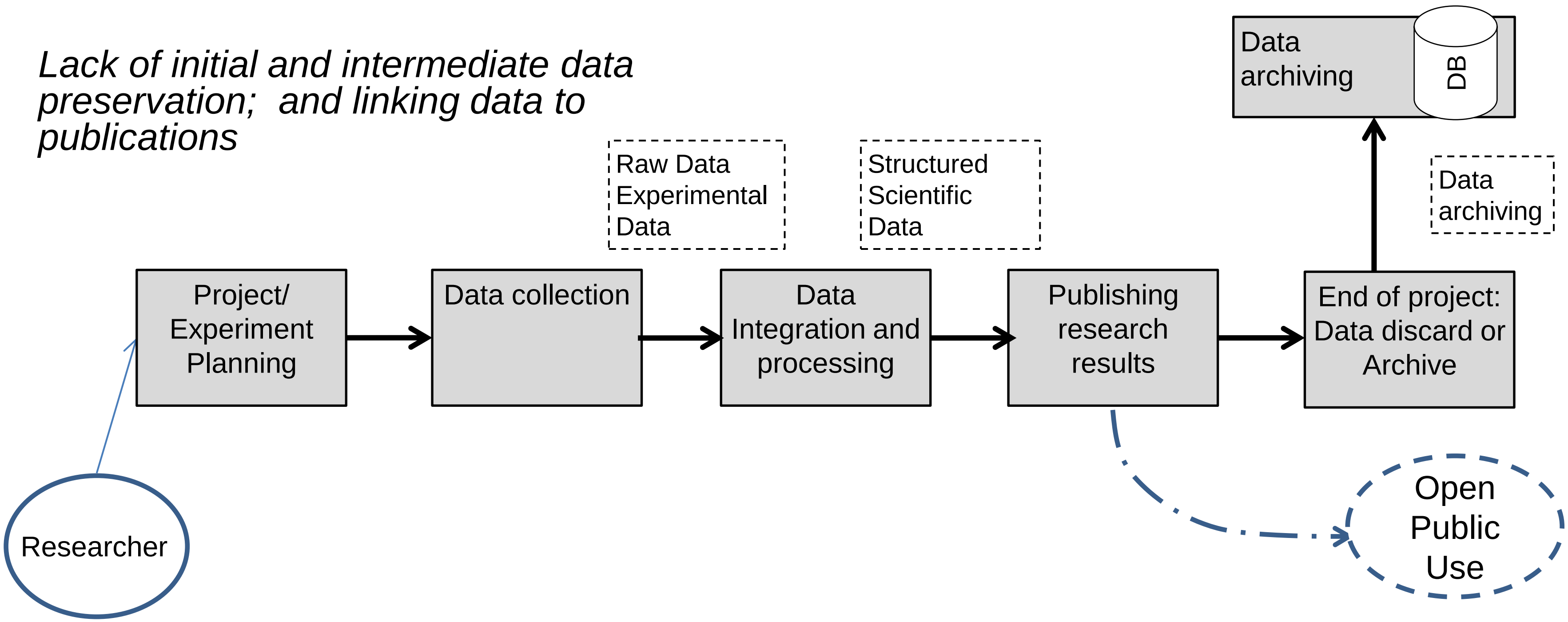


- **Raw data** collected from observation and from experiment (according to an initial research model)
- **Structured data** and datasets that went through data filtering and processing (supporting some particular formal model)
- **Published data** that supports one or another scientific hypothesis,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

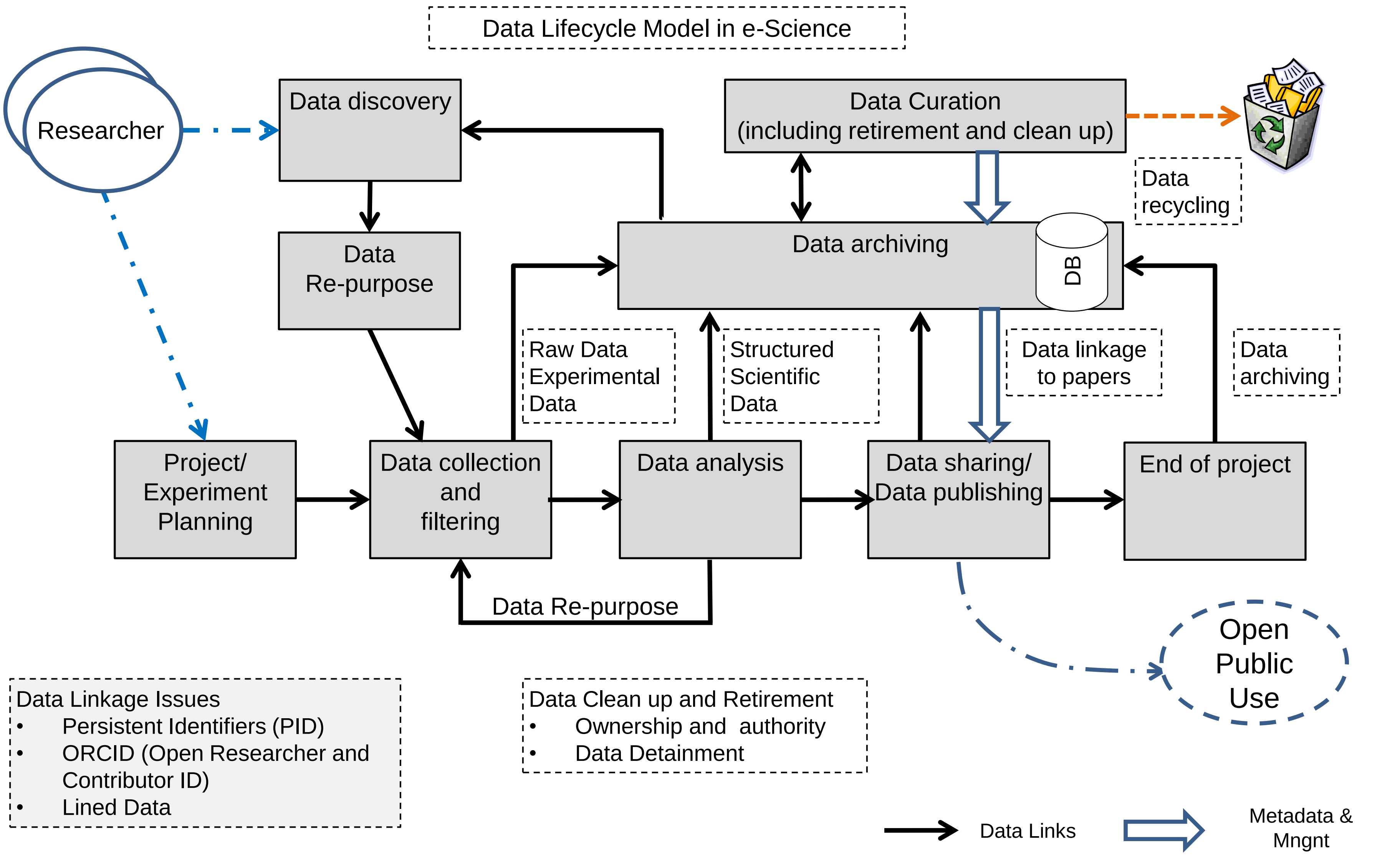
Traditional Data Lifecycle Model

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

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



Scientific Data Lifecycle Management (SDLM) Model



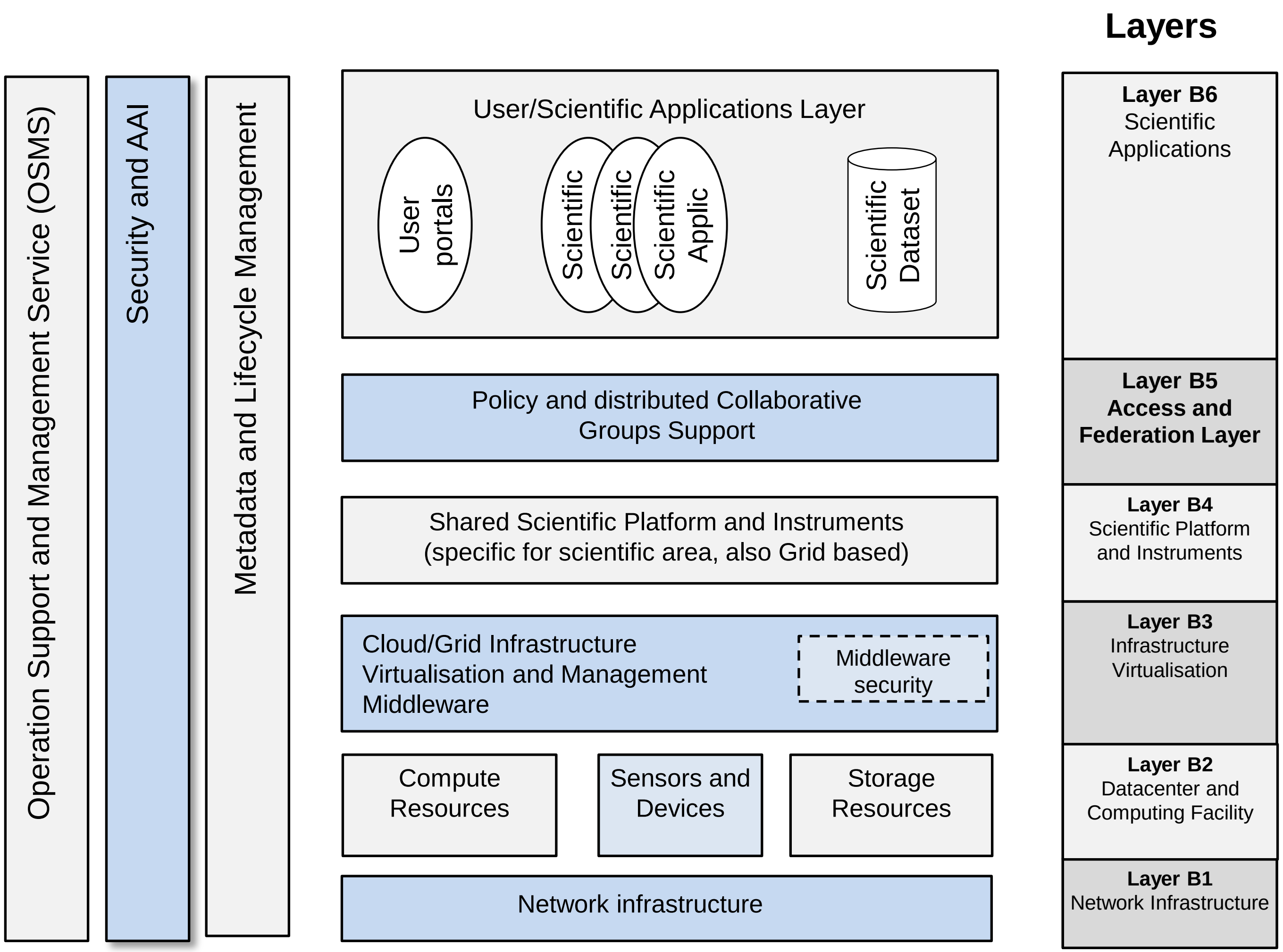
General requirements to SDI for emerging Big Data Science

- 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
  - Research need to trust SDI to put all their data on it
- 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



Technologies and solutions

Scientific specialist applications
Library resources

Optical Network Infrastructure
Federated Identity Management: eduGAIN, REFEDS, VOMS, InCommon

PRACE/DEISA

Grid/Cloud

Clouds

Autobahn, eduroam

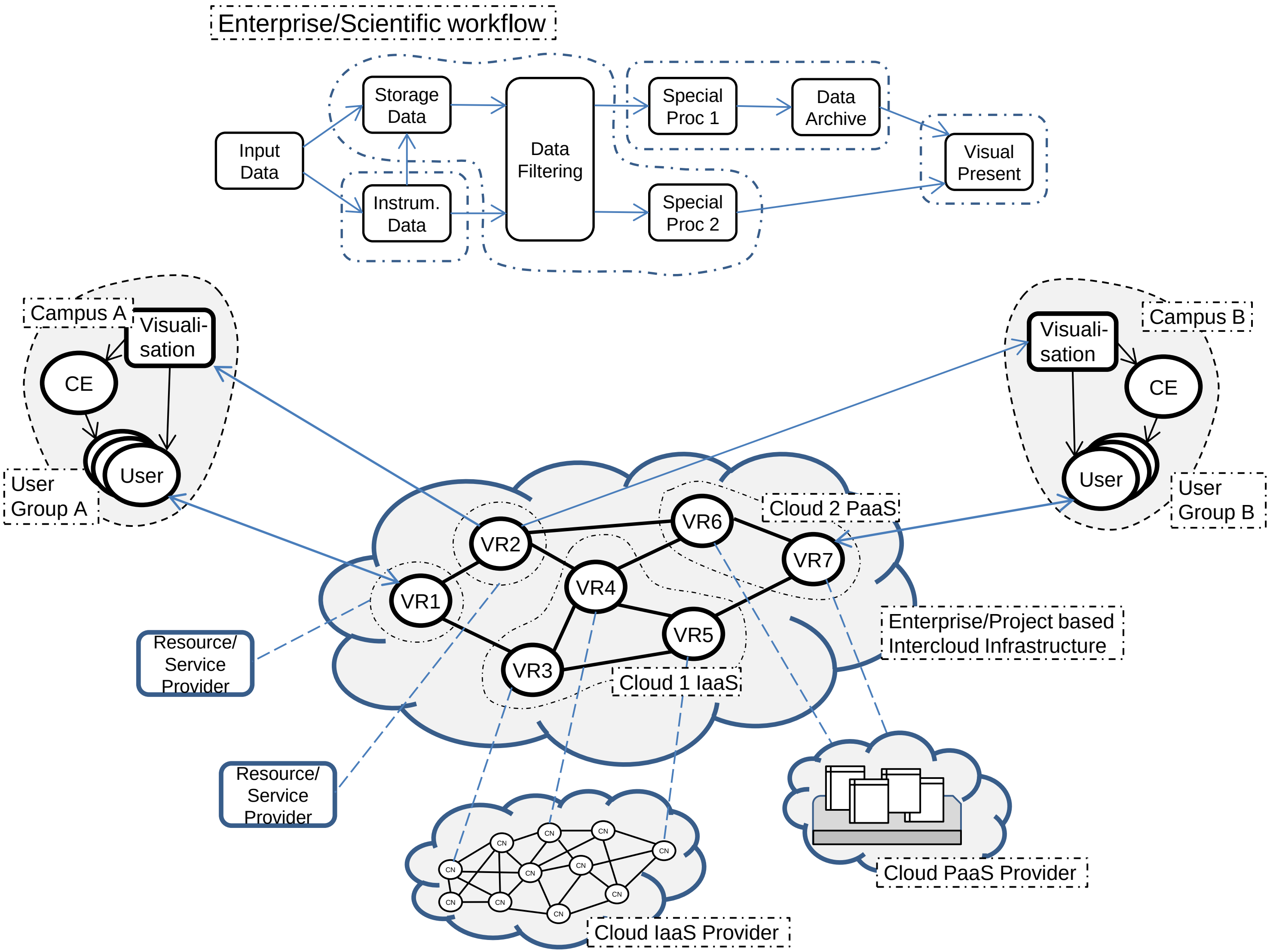
SDI Architecture Layers

- **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

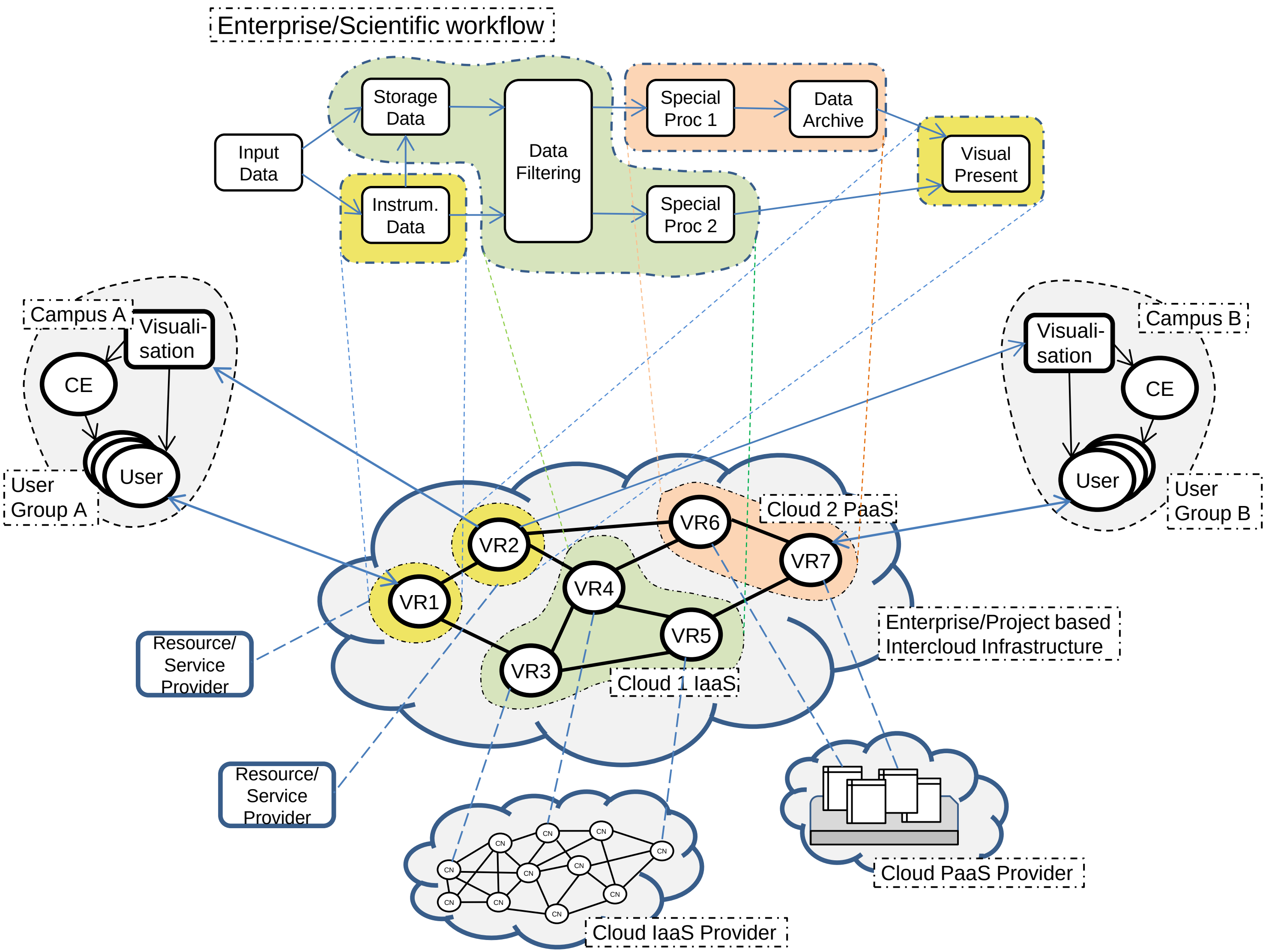
SDI move to Clouds

- 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
- Software Defined Infrastructure Services
  - As wider than currently emerging SDN (Software Defined Networks)
- Distributed Hadoop clusters for HPC and MPP

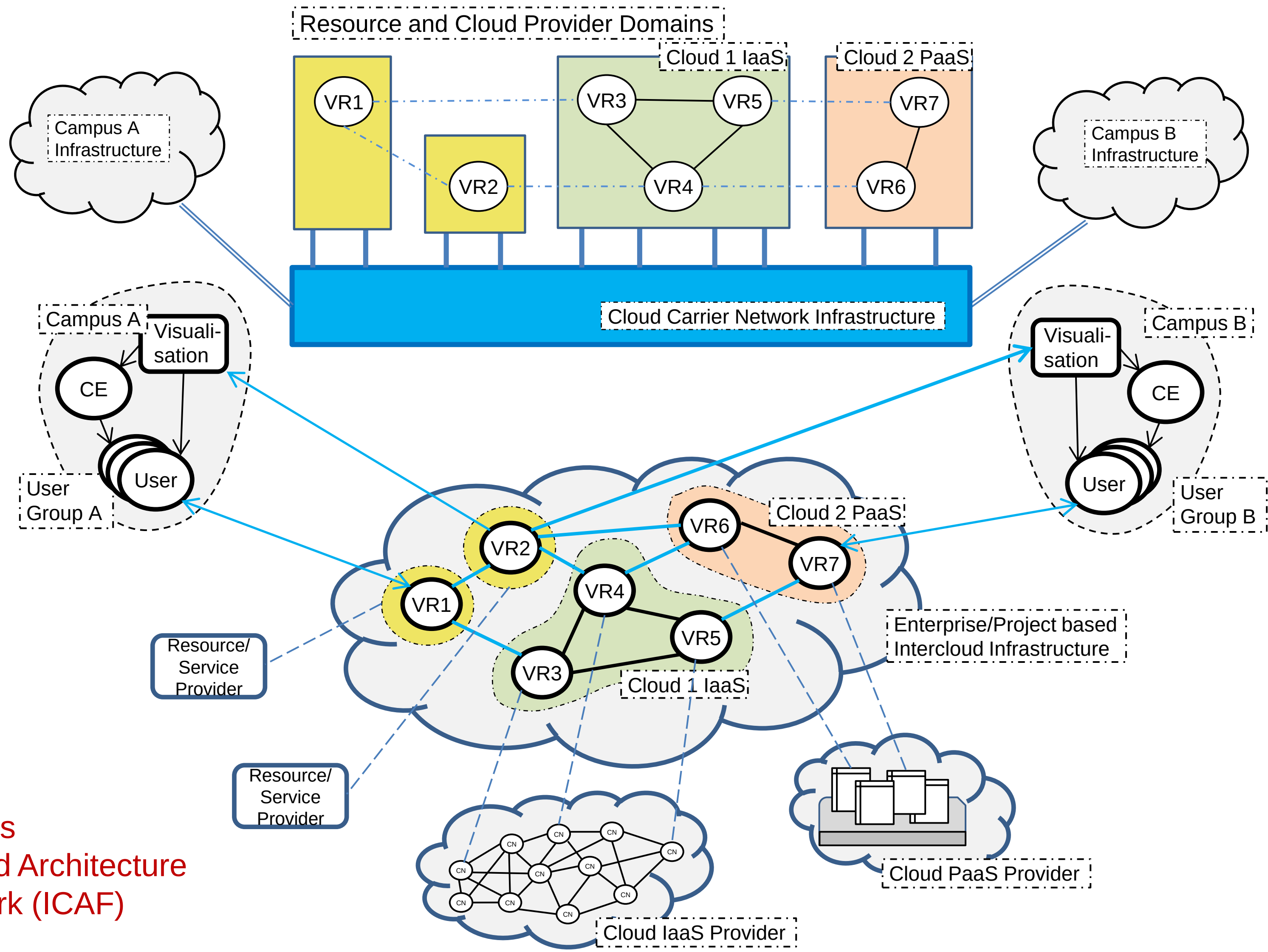
General use case for infrastructure provisioning: Workflow => Logical (Cloud) Infrastructure



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 Cloud/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)
  - Data centric security models
  - Authenticity, authorisation, delegation
  - Trust, trustworthy systems, validity
  - Accounting, auditing
  - Privacy: Will it change with the new technologies?

Questions and Discussion

Additional materials

- Big Data technologies and foreseen innovations
- Trust and trustworthiness

Big Data: Existing technologies (1)

- Big Data storage and database 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. optimised for big data, workflow and data structure

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 new more open technologies 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

Foreseen Big Data Innovations in 2013

- Cloud-Based Big Data Solutions
  - Amazon’s Elastic Map Reduce (EMR) is a market leader
  - Expected new innovative Big Data and Cloud solutions
- Real-Time Hadoop
  - Google’s Dremel-like solutions that will allow real-time queries on Big Data and be open source
- Distributed Machine Learning
  - Mahout iterative scalable distributed backpropagation machine learning and data mining algorithm
  - New algorithms Jubatus, HogWild
- 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
- 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/
- Big Data in 2013 by Mike Guattieri, Forrester
  - **Firms realize that Big Data means all it's data**. Big Data is the frontier of a firms ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and create better customer experiences. In 2013, all data is Big Data.
  - **The algorithm wars begin.** Advanced analytics uses advanced statistical, data mining, and machine learning algorithms to dig deeper to find patterns that you can’t see using traditional BI tools, simple queries, or rules.
  - **Real-time architectures will swing to prominence**. Firms will seek out streaming, event processing, and in-memory data technologies to provide real-time analytics and run predictive models. Mobile is a key driver because hyper-connected consumers and employees will require architectures that can quickly process incoming data from all digital channels to make business decisions and deliver engaging customer experiences in real-time.
  - **Big Data may be disruptive.** This is the year that Time magazine names Big Data as 2013 person of the year.

Trust model – Transposition to ICT/SDI domain

Classic Trust definition ("humanised")

A trusts B to do something

(typically Action, Operation, or Assertion)

Trust definition transposed to Trustworthy ICT

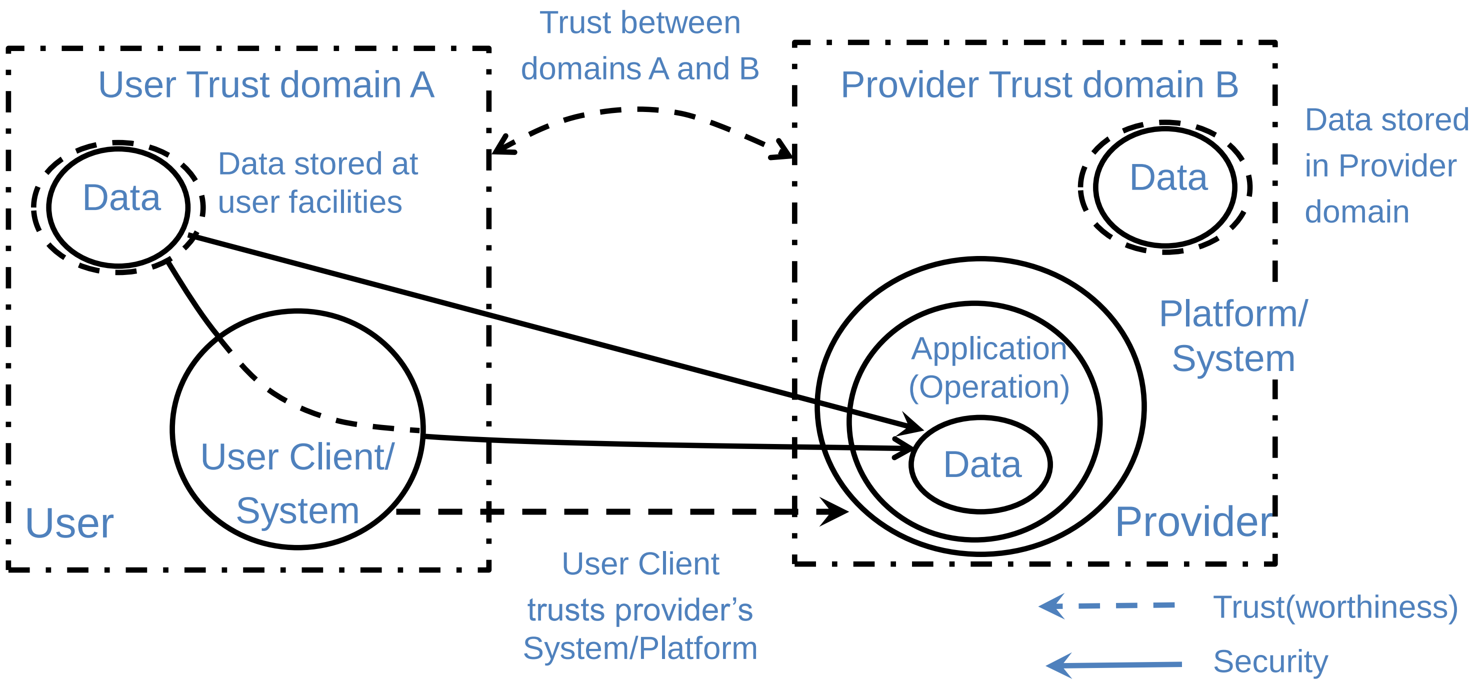
User A trusts service B to do an operation OP

(or make assertion)

with data D on a platform P

- Data are associated with the handling policy
- Platform is multi-tier, multi-domain and multi-provider
- Platform includes Computing, Storage and Network
- Data may be stored by 3rd party and belong to domain different to user

Component of the Trusted/Trustworthy ICT



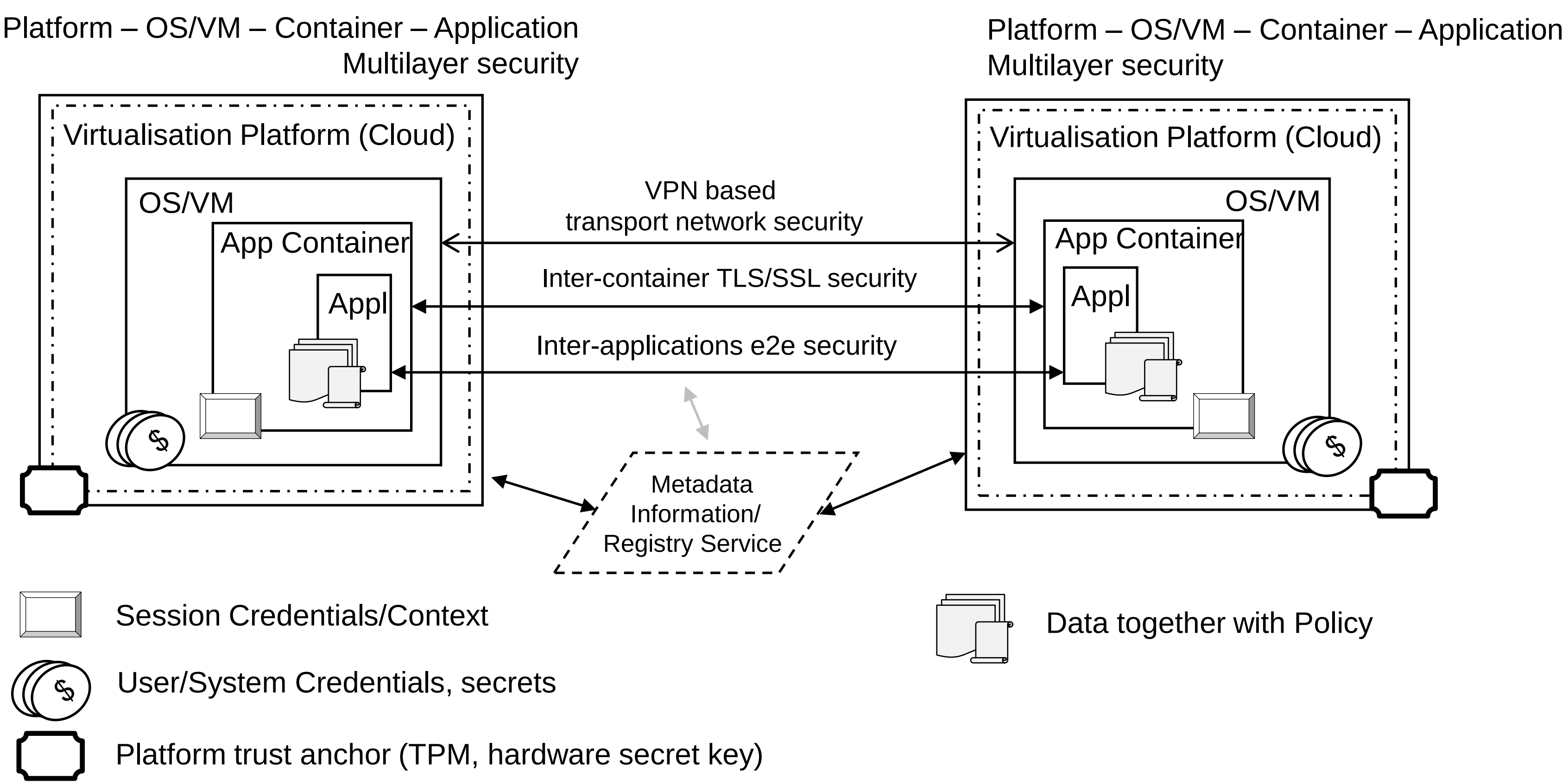
- User and Service Provider – the two actors concerned with own Data/Content security and each other System/Platform trustworthiness
- Data may be stored separately but typically belong to user trust domain In clouds it needs to be extended to distributed, multi-domain and multi-provider environment
- Two other aspects of security and trust
  - Data stored vs Data accessed/processed
  - System Idle vs Active (running User session)

Trustworthiness and Trustworthy system

Trustworthy infrastructure/system definition and analysis must include more factors

- Two or more parties involved into trust relations, typically Users (or user client/application) and Provider (or server or service/application)
- Data with their ownership, privacy, confidentiality level, and applied/attached handling policy
- Computing platform and storage (which however can associated with the general platform)
- Underlying communication infrastructure (which however can be associated with the general platform)
- Possible inter-domain issues related to different policies and assurance/provisions

Security in Virtualised Environment (2): Platform, OS/VM, Application and Network + Data



- Requires multilayer security and trust
  - Security and access control policies may be defined for each layer
- Data related components defining access control: metadata, encapsulated policies