BoF

Education and Skills Development in Data Intensive Science

RDA meeting 18-20 March 2013, Gothenburg
(Ljusvagen room)

(Yuri Demchenko, University of Amsterdam)

Outline

- Agenda bashing
- Round of introduction and interests expression

- Introduction to discussion (Yuri Demchenko, UvA)
- Any ad hoc discussions (TBD)

- Discussion on further steps

- BoF summary as reported to RDA Plenary meeting

BoF summary as reported to RDA Plenary

BoF on Education and Skills Development in Data Intensive Science

- Attended by 16 representatives from universities, libraries, e-Science, data centers, research coordination bodies
- Ad hoc agenda included
  - Round of introduction and interests expression
  - Introduction to discussion
    - H2020 priority area, existing approaches and experience in development of instructional materials for new technologies
  - Discussion on further steps
    - Priority topics to address and potential WG organisation

Topics discussed

- What experience do we have on component technologies to suport Scientific Data?
- Existing instructional and educational concepts and technologies
- How to benefit from collective knowledge and experience of RDA community?
- Scientific Data and Big Data in industry
  - Multidisciplinary domain and needs cooperation of specialists from multiple knowledge, scientific and technology domains

BoF Outcome and Decisions

- People agreed to work together both on defining joint activity and preparation of the next meeting
- Proceed with formal establishment of WG
- Involve LERU, LIBER and more US and overseas partners
- Hold next BoF with wider non-technical scope
  - Reach wider and targeted community and potential stakeholders and interested parties
  - Involve universities working on the DIS education programs
- Define output of potential WG addressing education and training on all aspects of (Scientific & Big) Data
  - Data management
  - Technology and infrastructure
  - Domain specific knowledge and user support
- Define skills/competencies vocabulary and Common Body of Knowledge for Data Intensive Science
  - Review existing/known education and training programs and frameworks

BoF Meeting slides and discussion

Possible discussion topics

- How to share experience between universities started programs development on Data Science?
  - What experience do we have on component technologies?
  - Existing instructional and educational concepts and technologies
- How to benefit from collective knowledge and experience of RDA community?
- TBD

Introduction to discussion on Education and Training for Data Intensive Science

- Big Data divide and need for Professional Education and Training
- Horizon2020 Priority area 7: Skills and new professions for research data
- Data Intensive Science and foundation technologies
  - Scientific Data Infrastructure (SDI) and Big Data Infrastructure (BDI)
- Example Cloud Computing Curriculum development
  - Cloud Computing as enabling technology for SDI and BDI
  - Cloud Computing Common Body of Knowledge
  - Course instructional approach: Bloom’s Taxonomy and Andragogy
  - Course structure Cloud Computing technologies and services design

Research Data e-Infrastructures: Framework for Action in Horizon2020

http://ec.europa.eu/digital-agenda/en/content/consultation-research-data-infrastructures-framework-action (comments deadline 27 March 2013)

e-Infrastructure fiches 01 – 07

01 - Community support data services

02 - Infrastructure for Open Access

03 - Storing, managing and preserving research data

04 - Discovery and provenance of research data

05 - Towards global data e-infrastructures

06 - Authentication and Authorisation e-infrastructures

07 - Skills and new professions for research data

https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/data_einfra_h2020_fiches_on-line_consult.pdf

e-Infrastructure fiche 07: Skills and new professions for research data

Scenario:

In a fast developing data-intensive world of scientific and scholarly research**, what kind of skills are needed for creating, handling, manipulating, analysing and storing for re-use of large amounts of data by others**? Some researchers in data-intensive research areas have acquired considerable skills in handling and managing data themselves or have a colleague who has these skills, but in many cases researchers turn to the institutional IT services or library for assistance and advice. Current data scientists usually end up in their roles accidentally as formal education hardly exists.

Actions address:

Defining or updating **university curricula** and **sharing best practices** across Europe; developing **training programmes for data scientists** working as part of a team of researchers or in close collaboration with them as responsible for computing facilities, storage and access; developing **training programmes for data librarians** from the library community who are specialised in the **curation, preservation and archiving of data**.

Stakeholders:

e-Infrastructure providers; software designers, education tools designers, associations such as LIBER, LERU, EAU, ScienceEurope, Universities, Libraries, etc.

Knowledge and/vs Skills in Data Intensive Science (DIS)

Target skills (knowledge?)

- Create, handle, manipulate, analyse and store large amounts of data for re-use of by others
- Develop, design and operate related data infrastructure shaped for specific projects, tasks, workflows

Currently expertise available with/from

- Practicing e-Scientists, institutional IT service specialists, or librarians dealing with data

Suggested Actions

- Defining or updating university curricula + defining training programs
- Sharing best practices across Europe

Developing training programmes for target specialist groups

- Data scientists who will work as part of a team of researchers or in close collaboration with them as responsible for computing facilities, storage and access
- Data librarians from the library community who will be specialised in the curation, preservation and archiving of data

Skills Vocabulary vs Common Body Knowledge

- Skills vocabulary describe what skills are identified
- CBK define what knowledge constitute professional knowledge in the area
- IEEE WG12: Learning Object Metadata (LOM) standards (2002-2004)
  - 1484.12.1: IEEE Standard For Learning Object Metadata
  - 1484.12.2 - 4: Binding For Learning Object Metadata Data for XML and RDF

Scientific (Big) Data Infrastructure definition

- Map/relate Big Data attributes/features to DFT/RDA terminology and Data Architecture/Model (by RDA?)

- Relation between Scientific (Big) Data and Big Data in industry

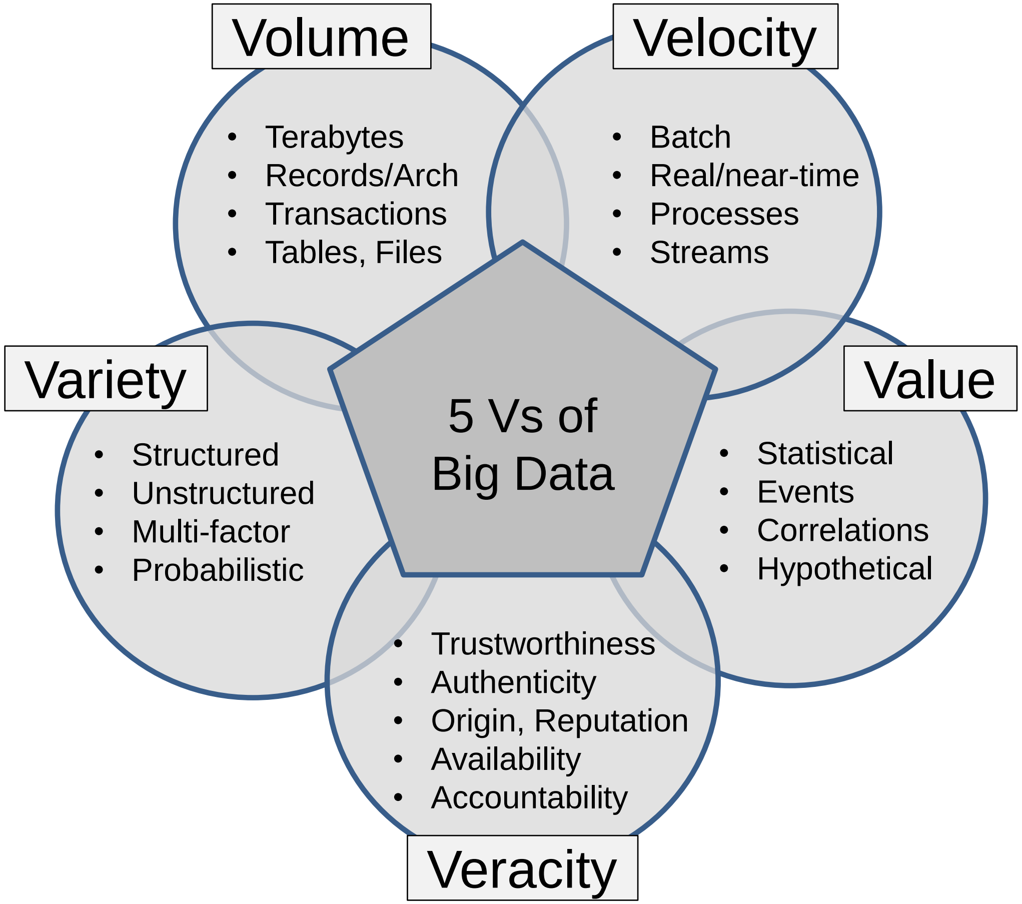
Big Data Definition

- 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

- 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"

5 V’s of Big Data



5 Vs of Big Data

Volume
- Terabytes
- Records/Arch
- Transactions
- Tables, Files

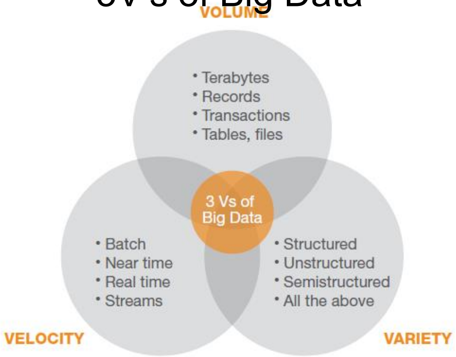
Velocity
- Batch
- Real/near-time
- Processes
- Streams

Variety
- Structured
- Unstructured
- Multi-factor
- Probabilistic

Value
- Statistical
- Events
- Correlations
- Hypothetical

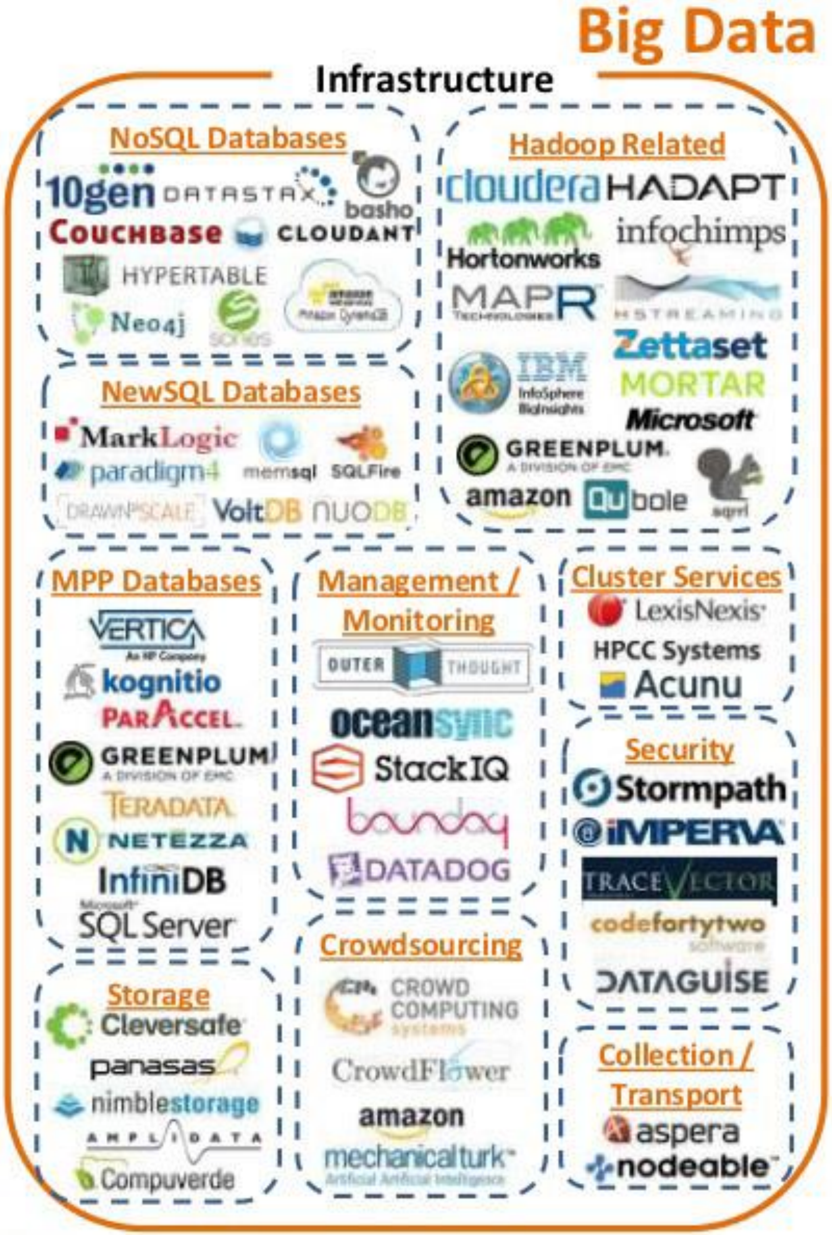
Veracity
- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability

Commonly accepted 3V’s of Big Data



Big Data Infrastructure Components

- Cloud base infrastructure services for data centric applications (storage, compute, infrastructure/VM management)
  - Software Defined Infrastructure
  - High performance switched network
- Hadoop/cluster related services and tools
- Specialised data analytics tools (logs, events, data mining, etc.)
- MPP (Massively Parallel Processing) applications (comoute&storage)
- Databases/Servers SQL, NoSQL
- Big Data Management
- Registries, Indexing/search, semantics, namespaces
- Security infrastructure (access control, policy, confidentiality, trust, privacy)
- Collaborative environment (federation, groups management)



Big Data Infrastructure Components

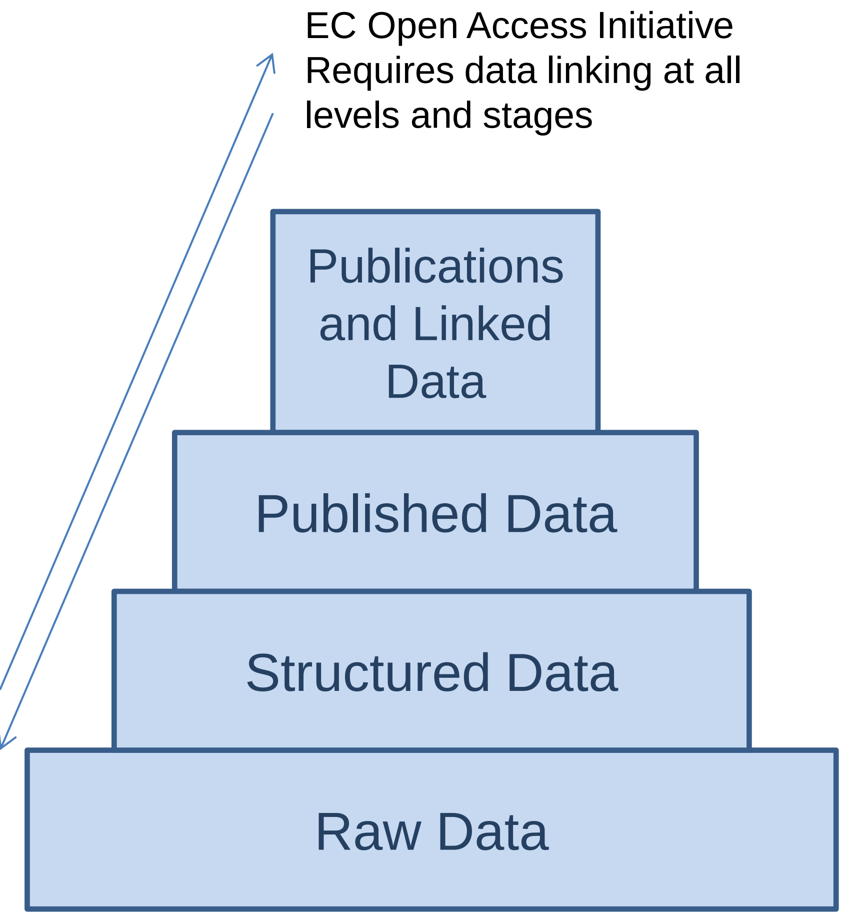
- Cloud base infrastructure services for data centric applications (storage, compute, infrastructure/VM management)
  - Software Defined Infrastructure
  - High performance switched network
- Hadoop/cluster related services and tools
- MPP (Massively Parallel Processing) applications (comoute&storage)
- Specialised data analytics tools (logs, events, data mining, etc.)
- Databases/Servers SQL, NoSQL
- Big Data Management
- Registries, Indexing/search, semantics, namespaces
- Security infrastructure (access control, policy, confidentiality, trust, privacy)
- Collaborative environment (federation, groups management)

Scientific Data Infrastructure Definition and Requirements

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

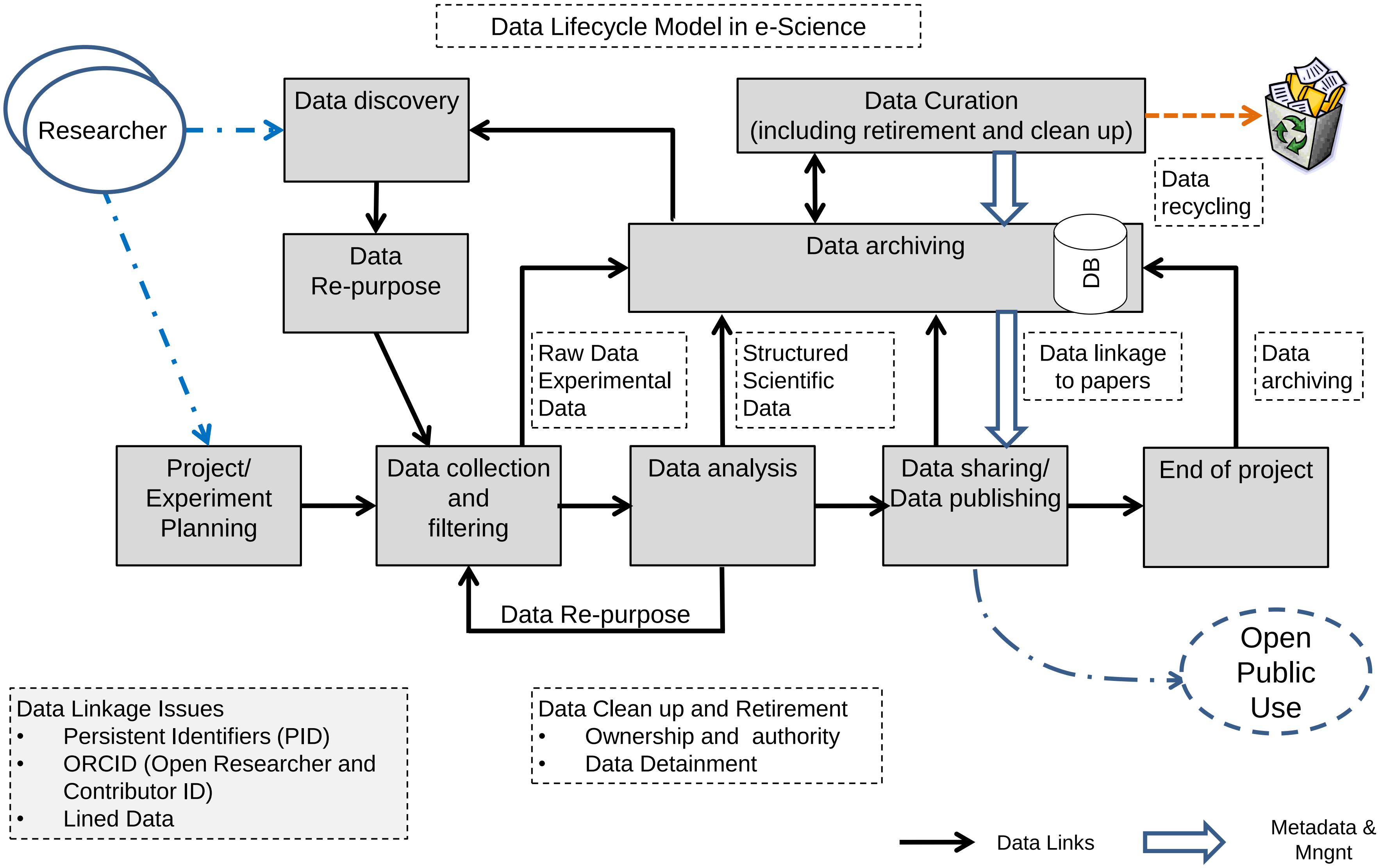
Published Data

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.

Scientific Data Lifecycle Management (SDLM) Model



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

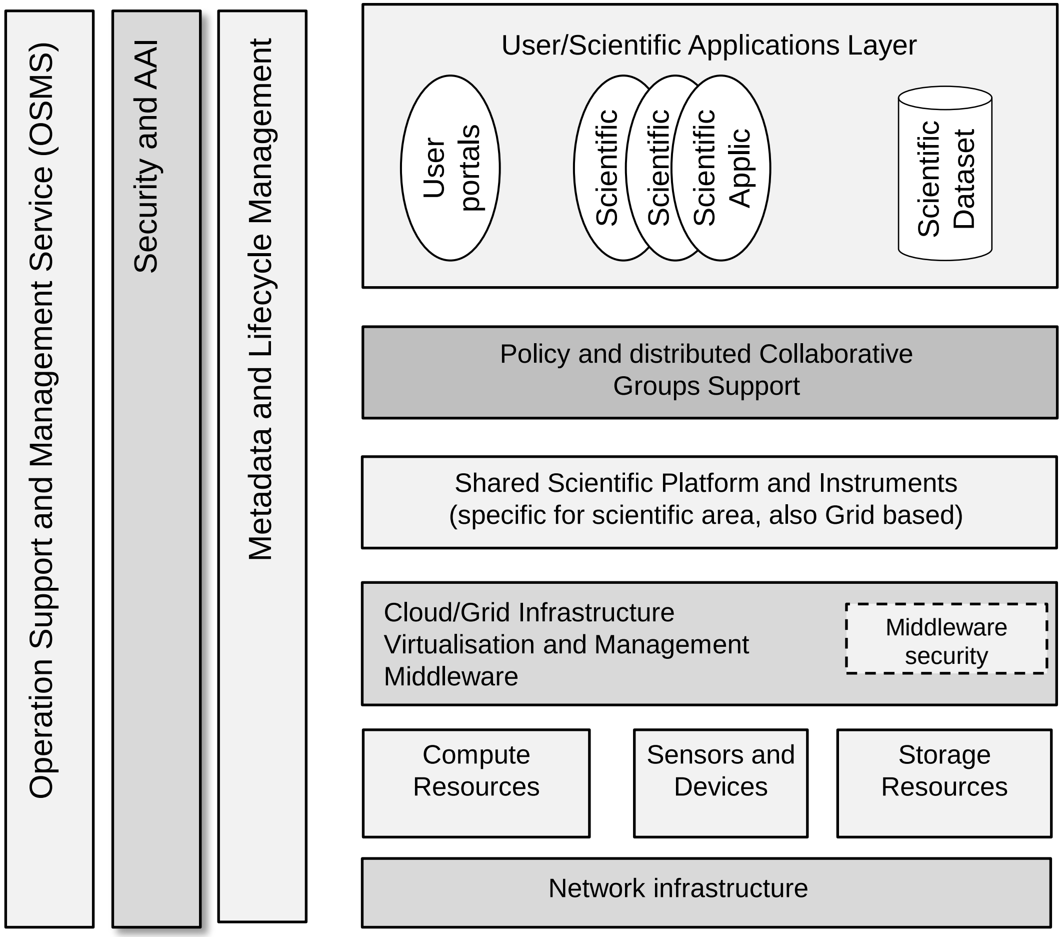
Data Links

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

SDI Architecture Model



Operation Support and Management Service (OSMS)

Security and AAI

Metadata and Lifecycle Management

User/Scientific Applications Layer

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	Technologies and solutions
Layer B6 Scientific Applications	Scientific specialist applications Library resources
Layer B5 **Access and Federation Layer**	Network Infrastructure eduGAIN, REFEDS, VOMS, InCommon
Layer B4 Scientific Platform and Instruments	PRACE/DEISA
Layer B3 Infrastructure Virtualisation	Grid/Cloud
Layer B2 Datacenter and Computing Facility	Clouds
Layer B1 Network Infrastructure	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

DIST Program development principles and example

- Reuse experience from the existing related program
  - Cloud Computing Technologies and Tools
  - Theoretical Informatics, Data Analytics and Artificial Intelligence
- General interactive education principles
  - Common Body of Knowledge
  - Bloom’s taxonomy
  - Pedagogy vs Andragogy
  - Discussion forum for online education, and Research and Reading assignments for on-campus eductaion

Example:
Cloud Computing Curriculum development

- Cloud Computing as enabling technology for Scientific Data Infrastructure (SDI)
- Cloud Computing Common Body of Knowledge
- Course instructional approach: Bloom’s Taxonomy and Andragogy
- Course structure Cloud Computing technologies and services design

Example: Common Body of Knowledge (CBK) in Cloud Computing

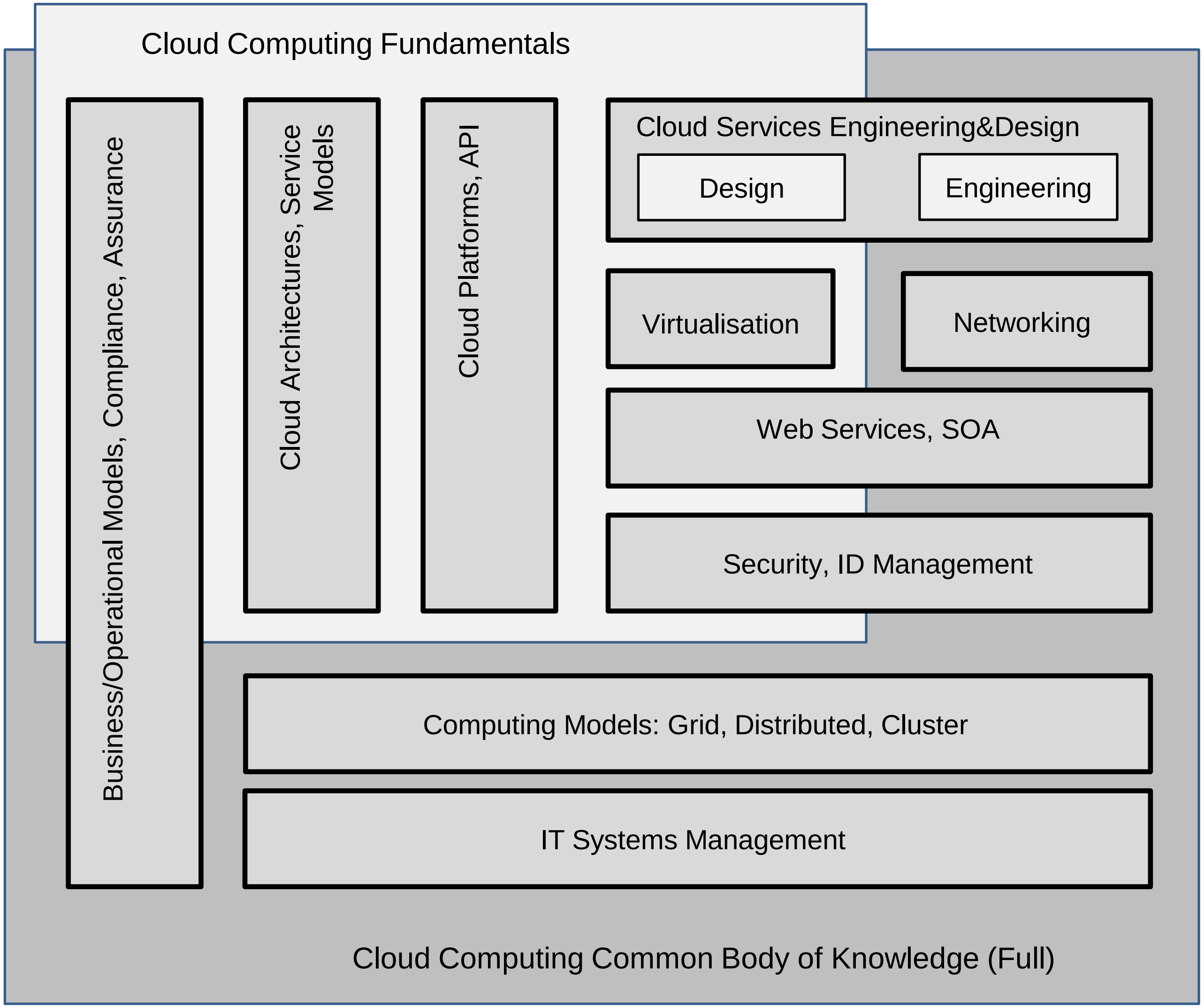
CBK refers to several domains or operational categories into which Cloud Computing theory and practices breaks down

- Still in development but already piloted by some companies, including industry certification program (e.g. IBM, AWS?)

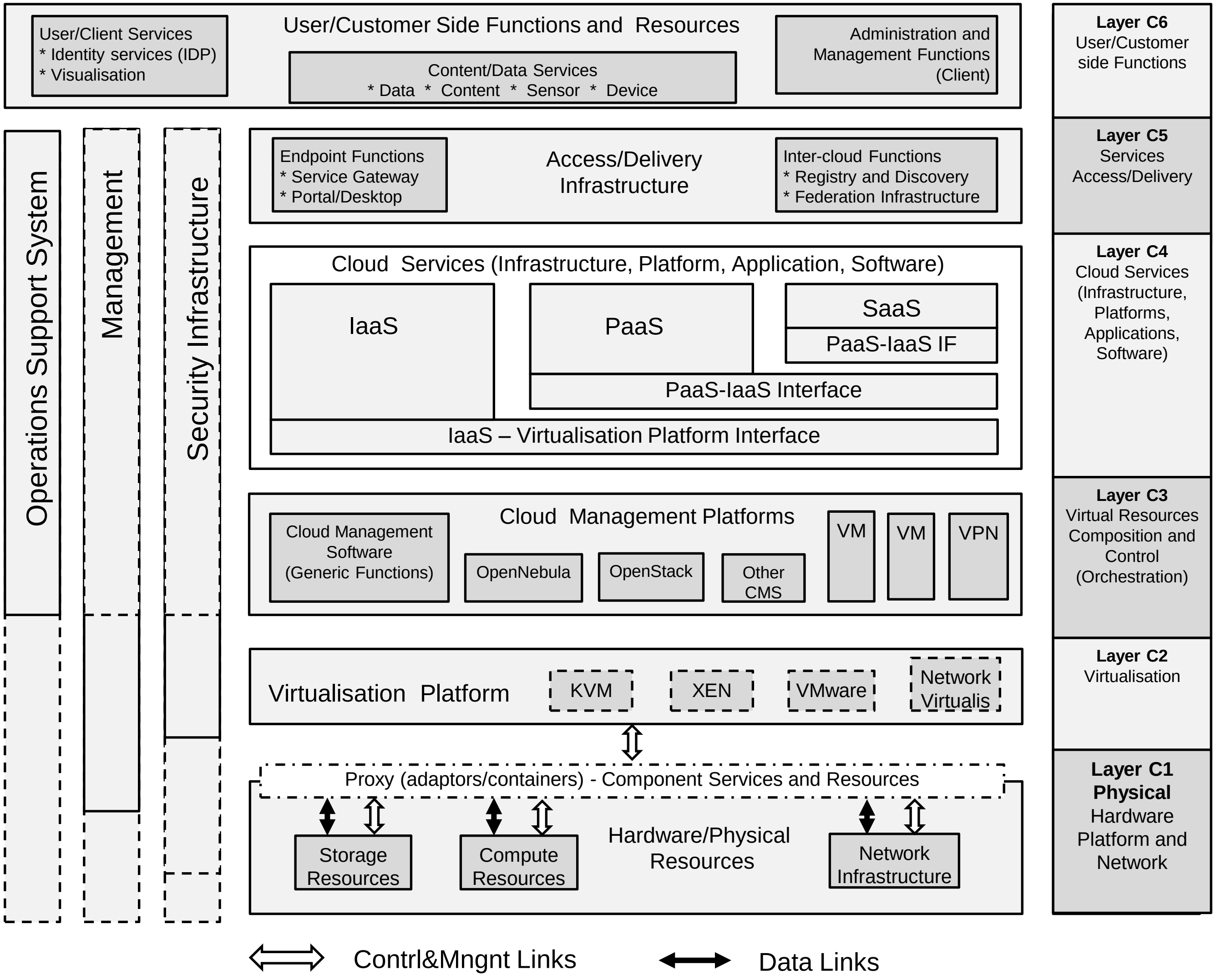
CBK Cloud Computing elements

1. ***Cloud Computing Architectures, service and deployment models***
2. ***Cloud Computing platforms, software/middleware and API's***
3. ***Cloud Services Engineering, Cloud aware Services Design***
4. Virtualisation technologies (Compute, Storage, Network)
5. Computer Networks, Software Defined Networks (SDN)
6. Service Computing, Web Services and Service Oriented Architecture (SOA)
7. Computing models: Grid, Distributed, Cluster Computing
8. Security Architecture and Models, Operational Security
9. IT Service Management, Business Continuity Planning (BCP)
10. Business and Operational Models, Compliance, Assurance, Certification

Example:
CKB-Cloud Components Landscape


Cloud Computing Fundamentals
Business/Operational Models, Compliance, Assurance
Cloud Architectures, Service Models
Cloud Platforms, API
Cloud Services Engineering&Design
Design
Engineering
Virtualisation
Networking
Web Services, SOA
Security, ID Management
Computing Models: Grid, Distributed, Cluster
IT Systems Management
Cloud Computing Common Body of Knowledge (Full)

Example: Multilayer Cloud Services Model (CSM) – Taxonomy of Existing Cloud Architecture Models


User/Customer Side Functions and Resources
User/Client Services
* Identity services (IDP)
* Visualisation
Content/Data Services
* Data * Content * Sensor * Device
Administration and Management Functions (Client)
Operations Support System
Management
Security Infrastructure
Endpoint Functions
* Service Gateway
* Portal/Desktop
Access/Delivery Infrastructure
Inter-cloud Functions
* Registry and Discovery
* Federation Infrastructure
Cloud Services (Infrastructure, Platform, Application, Software)
IaaS
PaaS
SaaS
PaaS-IaaS IF
PaaS-IaaS Interface
IaaS – Virtualisation Platform Interface
Cloud Management Platforms
Cloud Management Software (Generic Functions)
OpenNebula
OpenStack
Other CMS
VM
VM
VPN
Virtualisation Platform
KVM
XEN
VMware
Network Virtualis
Proxy (adaptors/containers) - Component Services and Resources
Storage Resources
Compute Resources
Hardware/Physical Resources
Network Infrastructure
Contrl&Mngnt Links
Data Links
Layer C6
User/Customer side Functions
Layer C5
Services Access/Delivery
Layer C4
Cloud Services (Infrastructure, Platform, Applications, Software)
Layer C3
Virtual Resources Composition and Control (Orchestration)
Layer C2
Virtualisation
Layer C1
Physical
Hardware Platform and Network

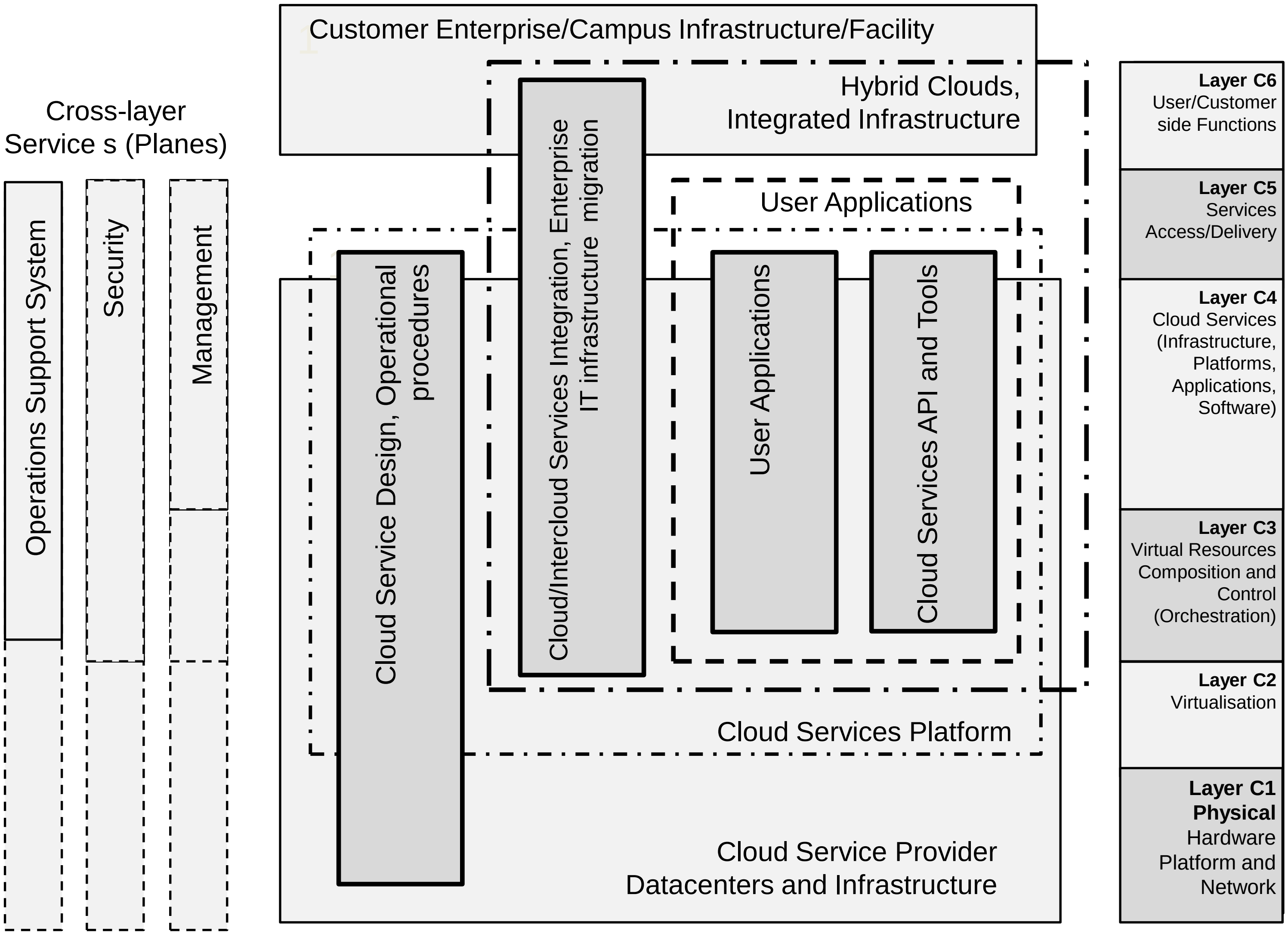
CSM layers

(C6) User/Customer side Functions
(C5) Services Access/Delivery
(C4) Cloud Services (Infrastructure, Platform, Applications)
(C3) Virtual Resources Composition and Orchestration
(C2) Virtualisation Layer
(C1) Hardware platform and dedicated network infrastructure

⇔ Control/ Mngnt Links

⟷ Data Links

Example:
Relations Course Components and CSM


Customer Enterprise/Campus Infrastructure/Facility
Hybrid Clouds,
Integrated Infrastructure
Cross-layer
Service s (Planes)
Operations Support System
Security
Management
Cloud Service Design, Operational procedures
Cloud/Intercloud Services Integration, Enterprise IT infrastructure migration
User Applications
User Applications
Cloud Services API and Tools
Cloud Services Platform
Cloud Service Provider
Datacenters and Infrastructure
Layer C6
User/Customer side Functions
Layer C5
Services Access/Delivery
Layer C4
Cloud Services (Infrastructure, Platforms, Applications, Software)
Layer C3
Virtual Resources Composition and Control (Orchestration)
Layer C2
Virtualisation
Layer C1
Physical
Hardware Platform and Network

Example:
Professional Knowledge in Cloud Computing

- (General) Professional level of knowledge includes but not limited to
  - Knowing basic concepts and major application areas
  - Knowing similar concepts (and concepts inter-relation) and alternatives, as well as application specific areas
  - Knowing basic technologies and their relation to basic concepts
  - Knowing authoritative (and not authoritative) sources of information and how to evaluate quality of information
    - Ability to work with standards (what is not an easy source of information)
    - Ability to critically evaluate and filter some inconsistent information, e.g. popular sites like wikipedia and similar, blogs, etc.
    - Critically evaluate vendors' information which is sometimes biased and/or doesn't provide enough background information

- Cloud computing is a new technology but it is becoming a common preferred base/platform for all current and future developments
- Becoming an expert in Cloud Computing
  - General professional knowledge and understanding of the main development areas
  - Practical development and experience with few projects, writing reports, technical documents, *following and contributing to standardisation*
  - Cloud aware and cloud powered analysis and thinking

Example: Professional Education in Cloud Computing - Principles

- Provide knowledge both in **Cloud Computing** as a new technology and **background technologies**
- Empower the future professionals with ability to **develop new knowledge** and build stronger expertise, prepare basis for new **emerging technologies** such as **Big Data**
- **Bloom's Taxonomy** as a basis for defining learning targets and modules outcome
  - Provides a basis for knowledge testing and certification
- **Andragogy vs Pedagogy** as instructional methodology for professional education and training
  - Course format: On-campus education and training, online courses, self-study

Example:
Bloom’s Taxonomy – Cognitive Activities

Knowledge

Exhibit memory of previously learned materials by recalling facts, terms, basic concepts and answers

- Knowledge of specifics - terminology, specific facts
- Knowledge of ways and means of dealing with specifics - conventions, trends and sequences, classifications and categories, criteria, methodology
- Knowledge of the universals and abstractions in a field - principles and generalizations, theories and structures
- Questions like: What are the main benefits of outsourcing company’s IT services to cloud?

Comprehension

Demonstrate understanding of facts and ideas by organizing, comparing, translating, interpreting, describing, and stating the main ideas

- Translation, Interpretation, Extrapolation
- Questions like: Compare the business and operational models of private clouds and hybrid clouds.

Application

Using new knowledge. Solve problems in new situations by applying acquired knowledge, facts, techniques and rules in a different way

- Questions like: Which cloud service model is best suited for medium size software development company, and why?

Analysis

Examine and break information into parts by identifying motives or causes. Make inferences and find evidence to support generalizations

- Analysis of elements, relationships, organizational principles
- Questions like: What cloud services are needed to support typical business processes of a web trading company, give suggestions how these services can be implemented with PaaS or IaaS clouds. Provide references to support your statements.

Synthesis

Compile information together in a different way by combining elements in a new pattern or proposing alternative solutions

- Production of a unique communication, a plan, or proposed set of operations, derivation of a set of abstract relations
- Questions like: Describe the main steps and tasks for migrating IT services of an example company to clouds, what services and data can be moved to clouds and which will remain at the enterprise premises.

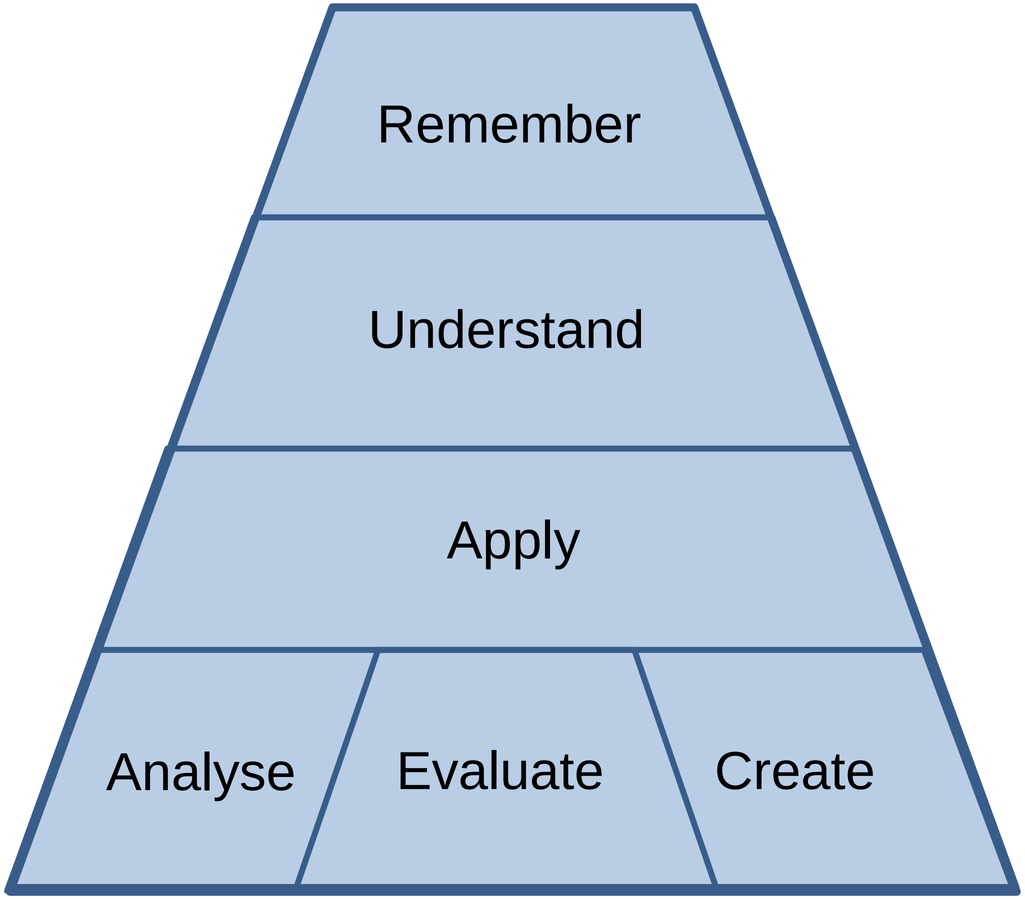
Evaluation

Present and defend opinions by making judgments about information, validity of ideas or quality of work based on a set of criteria

- Judgments in terms of internal evidence or external criteria
- Questions like: Do you think that cloudification of the enterprise infrastructure creates benefits for enterprises, short term and long term?

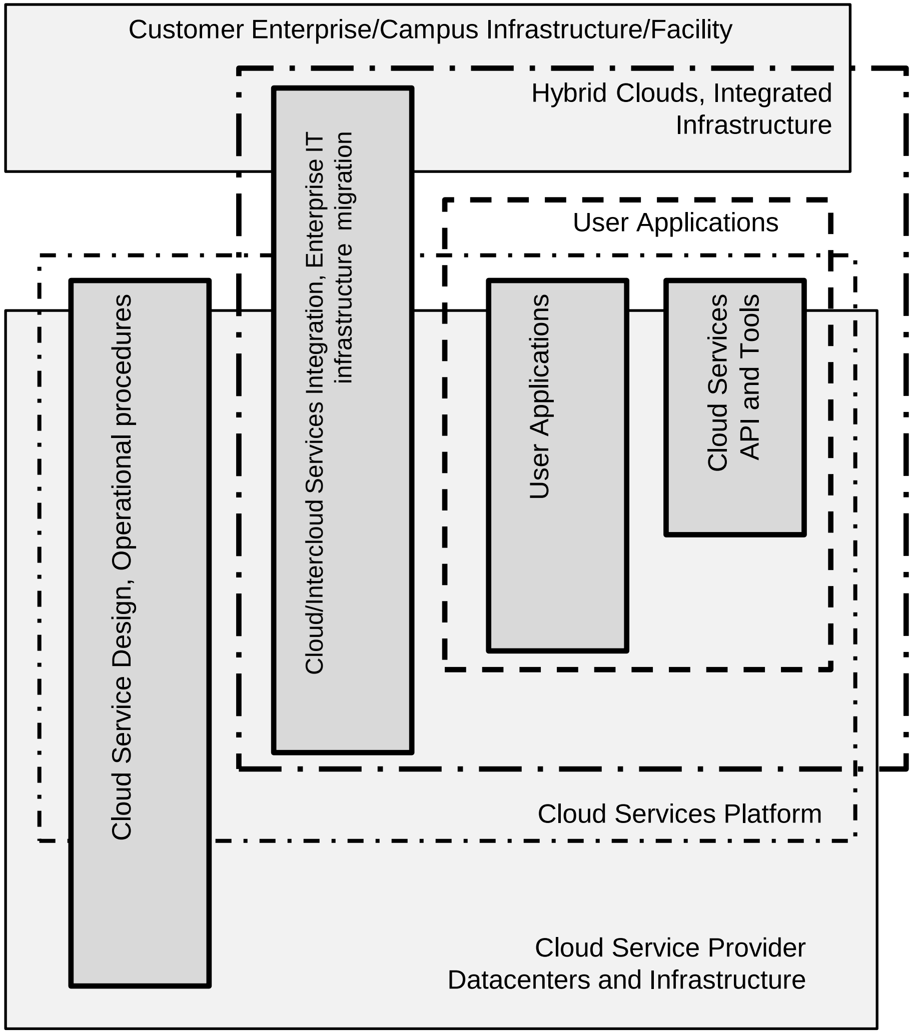
Example: Mapping Bloom's Taxonomy from Cognitive Domain to Professional Activity Domain

- Perform standard tasks, use API and Guidelines
- Create own complex applications using standard API (simple engineering)
- Integrate different systems/components, e.g. Cloud provider and enterprise (complex engineering)
- Extend existing services, design new services
- Develop new architecture and models, platforms and infrastructures



Example: Mapping Course Components, Cloud Professional Activity and Bloom’s Taxonomy

Taxonomy Cognitive Domain [3]	Taxonomy Professional Activity Domain
Knowledge	Perform standard tasks, use standard API and Guidelines
Comprehension	Create own complex applications using standard API (simple engineering)
Application	Integrate different systems/components, e.g. provider and enterprise infrastructure
Analysis	Extend existing services, design new services
Synthesis	Develop new architecture and models, platforms and infrastructures
Evaluation	


Customer Enterprise/Campus Infrastructure/Facility
Hybrid Clouds, Integrated Infrastructure
Cloud/Intercloud Services Integration, Enterprise IT infrastructure migration
User Applications
User Applications
Cloud Services API and Tools
Cloud Service Design, Operational procedures
Cloud Services Platform
Cloud Service Provider Datacenters and Infrastructure

Example: Pedagogy vs Andragogy

Pedagogy (child-leading) and Andragogy (man-leading)

- On-campus and on-line education
- Developed by American educator Malcolm Knowles, stated with six assumptions related to motivation of adult learning:
  - Adults need to know the reason for learning something (Need to Know)
  - Experience (including error) provides the basis for learning activities (Foundation)
  - Adults need to be responsible for their decisions on education; involvement in the planning and evaluation of their instruction (Self-concept)
  - Adults are most interested in learning subjects having immediate relevance to their work and/or personal lives (Readiness)
  - Adult learning is problem-centered rather than content-oriented (Orientation)
  - Adults respond better to internal versus external motivators (Motivation)

Example: Applying Andragogy to Self-Education and Online Training - Problems

- Andragogy concept is widely used in on-line education. However
  - Based on active discussion activities guided/moderated by instructor/moderator
  - Combined with the Bloom’s taxonomy
- Self-education (guided) and online training specifics
  - Course consistency in sense of style, presentation/graphics, etc.
  - Requires the course workflow to be maximum automated
    - Especially if coupled with certification or pre-certification
  - Less time to be devoted by trainee
    - Estimated 1 hour per lesson, maximum 3 lessons per topic
  - Knowledge control questionnaires at the end of lessons or topics

Example Cloud Computing Course Structure

Part 1.1. Cloud Computing definition and general usecases

Part 1.2. Cloud Computing and enabling technologies

Part 2.1. Cloud Architecture models and industry standardisation: Architectures

Part 2.2. Cloud Architecture models and industry standardisation: Standard interfaces

Part 3.1. Major cloud provider platforms

Part 3.2. Major cloud provider platforms: Research and Community Clouds

Part 4. Cloud middleware platforms (architecture, API, usage examples)

Part 5.1. Cloud Infrastructure as a Service (IaaS): Architecture, platform and providers

Part 5.2. Cloud Infrastructure as a Service (IaaS): IaaS services design and management

Part 6.1. Cloud Platform as a Service (PaaS): Architecture, platform and providers

Part 6.2. Cloud Platform as a Service (PaaS): PaaS services design and management

Part 7.1. Security issues and practices in clouds

Part 7.2. Security services design in clouds; security models and Identity management

Part 8 (Advanced). InterCloud Architecture Framework (ICAF) for Interoperability and Integration: Architecture definition and design patterns

Basic parts & Advanced parts

Example: Defining Target Audience

- Basic profile ("essentials" or “fundamentals”) is for IT decision makers, informed users:
  - all concepts are explained, clouds opportunities are demonstrated, general use cases are analysed, examples of use are provided
  - general security issues in clouds are explained
  - intended the course will allow this group of listeners to be able to understand what they need to learn more.
- "Advanced" part is for engineers/practitioners developing cloud services, and doing integration/consulting work:
  - different cloud architectures are explained, details on the different cloud related open interfaces (like CDMI, OVF, OCCI) and proprietary API (like Amazon AWS API) are provided, detailed overview of popular cloud platform/middleware (like OpenStack, OpenNebula, Eucalyptus) is provided
  - security models (including main cloud providers) and technologies explained, federated cloud identity and access control is explained
  - provide advice/suggestions where to look for further information