Developing Customisable Education and Training Program on Cloud Computing and Approach to Big Data Education

Yuri Demchenko
SNE Group, University of Amsterdam

BoF “Cloud Computing and Data Analysis Training for the Developing World”
18 September 2013, 2nd RDA Plenary

Outline

- Cloud Computing Curriculum design: Basic principles
- Proposed Big Data Architecture Framework (BDAF)
  - Data Models and Big Data Lifecycle
- RDA2 BoF on Educations and Skills Development for Data Intensive Science (17 Sept 2013)
- Discussion: Opportunities to support training in developing world (DW)

Cloud Computing Curriculum Development

- Presented at the BoF during the 1st RDA meeting 18-20 March 2013 in Gothenburg http://www.uazone.org/demch/presentations/rda2013-gothenburg-bof-education-skills-v02.pdf
- Cloud Computing as enabling technology for Scientific Data Infrastructure (SDI) and Big Data Infrastructure
- Cloud Computing Common Body of Knowledge (CBK)
- 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

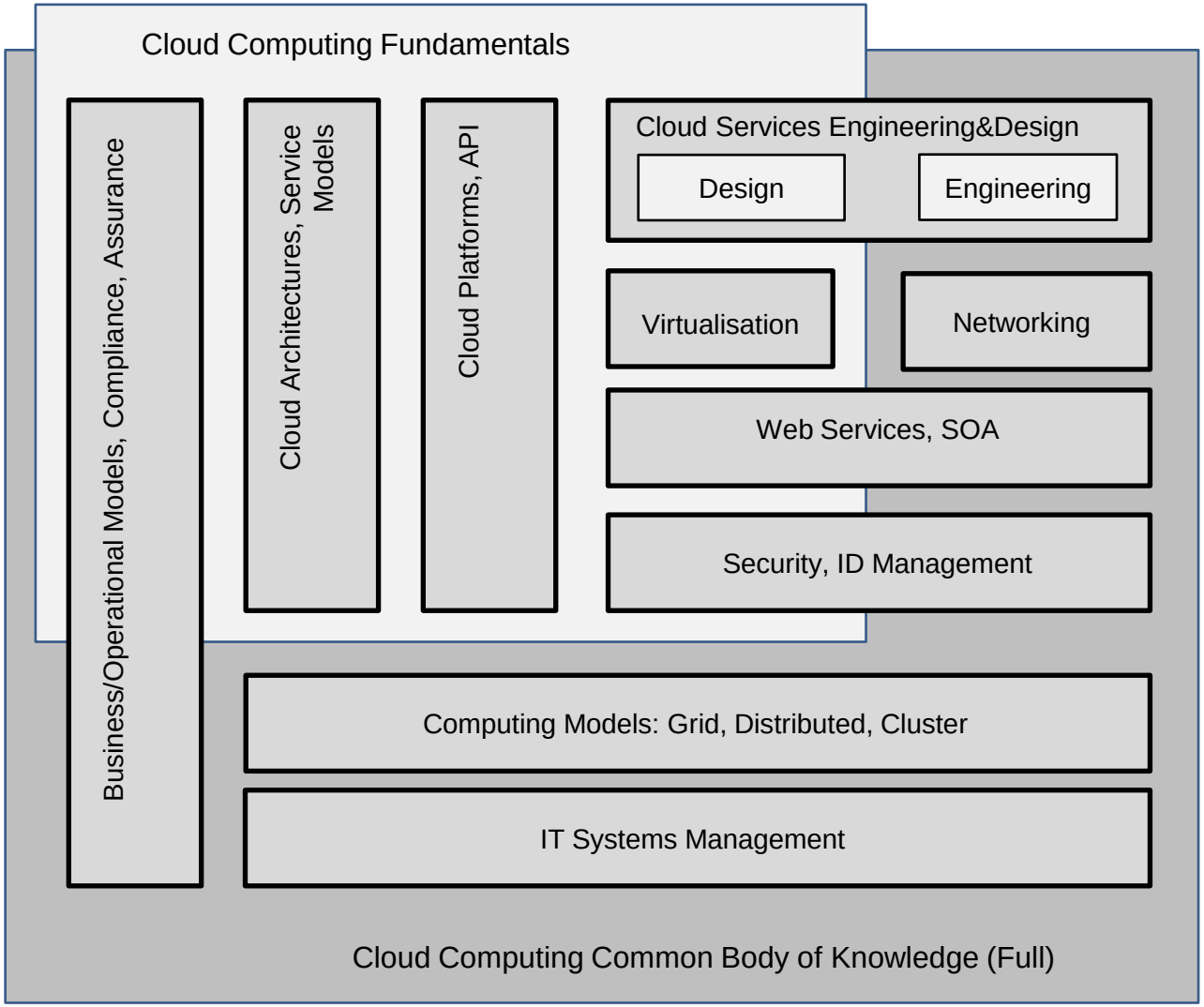
CBK refers to several domains or operational categories into which Cloud Computing theory and practices breaks down

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

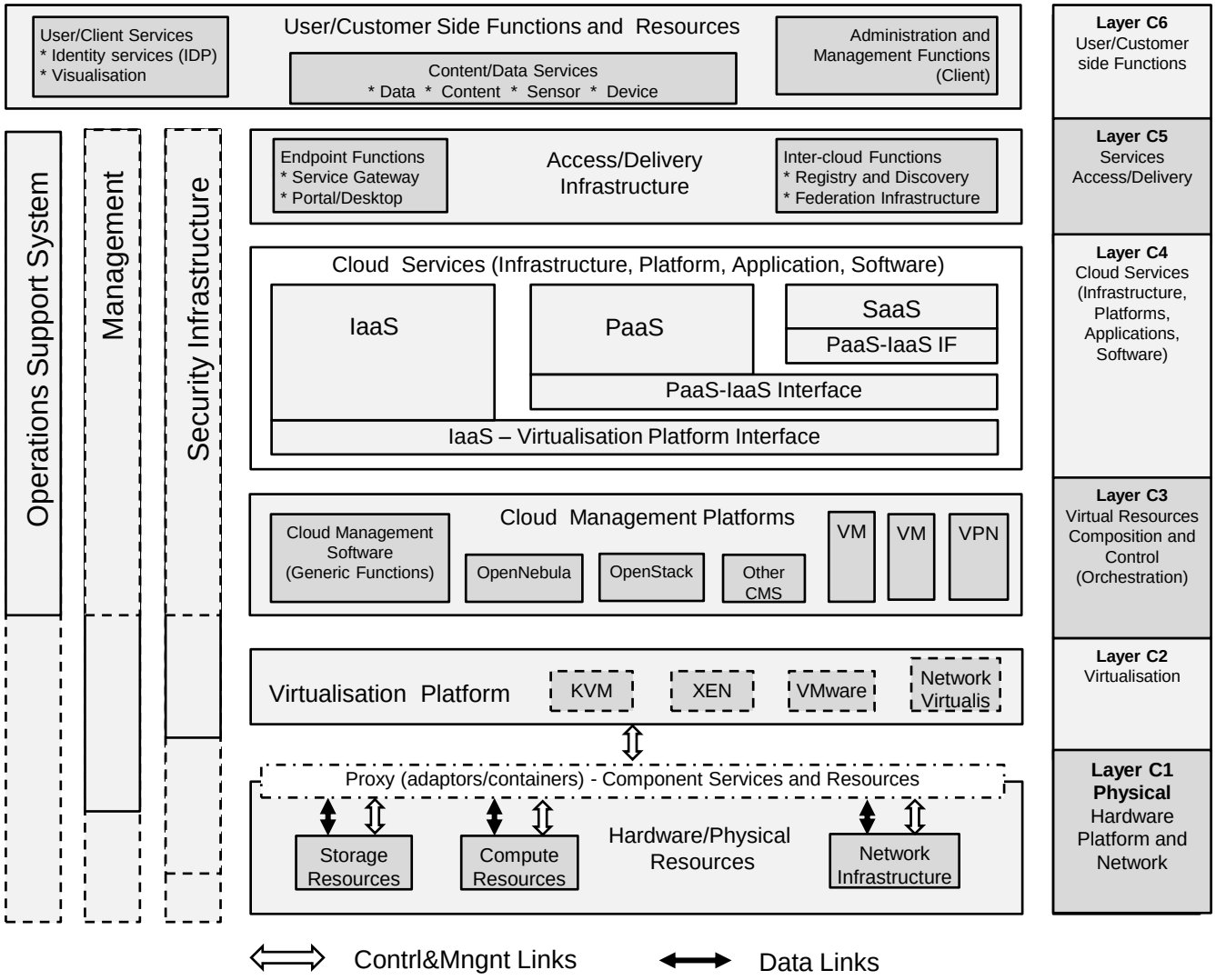
CBK Cloud Computing elements

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



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



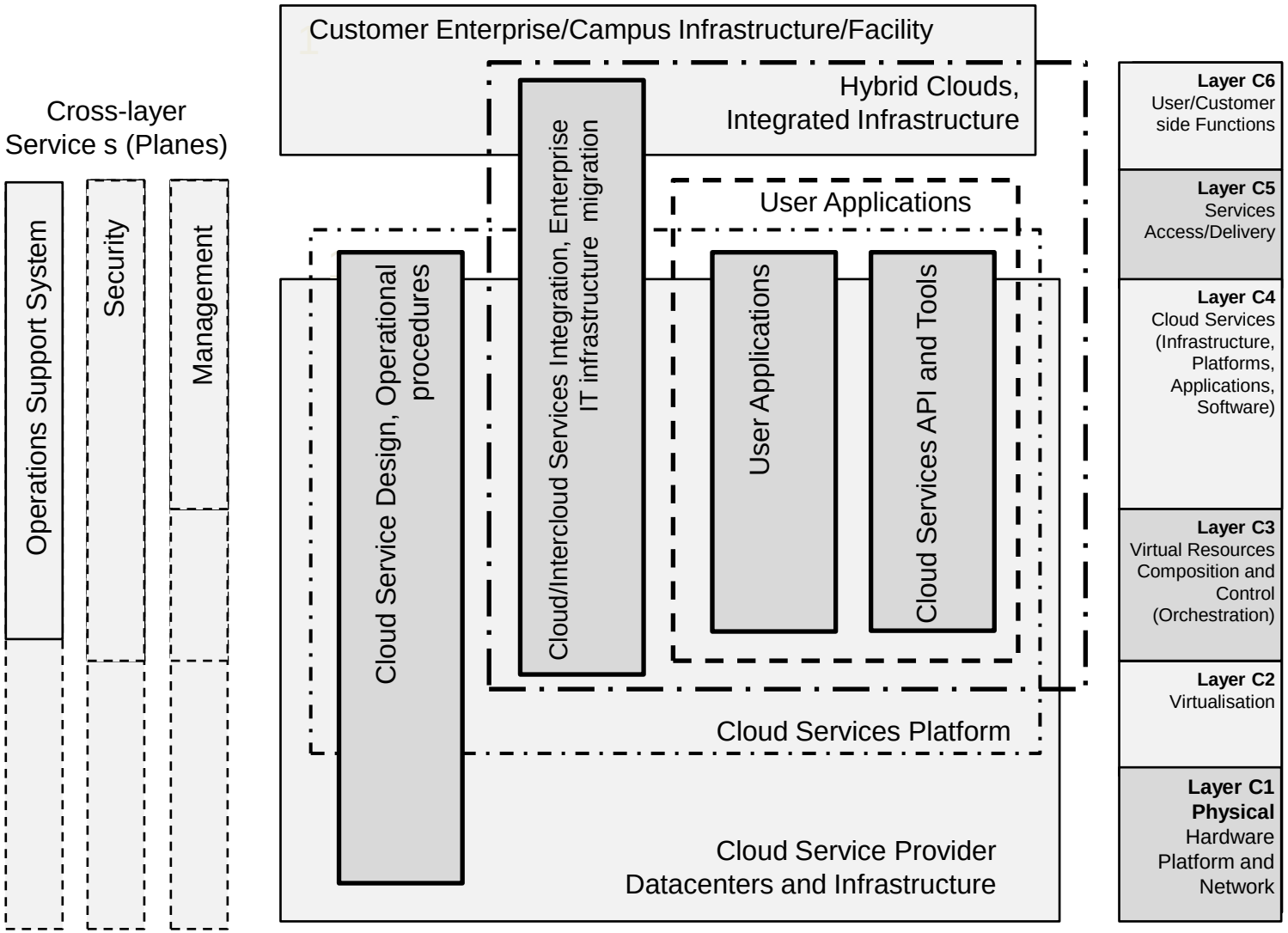
CSM layers

(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

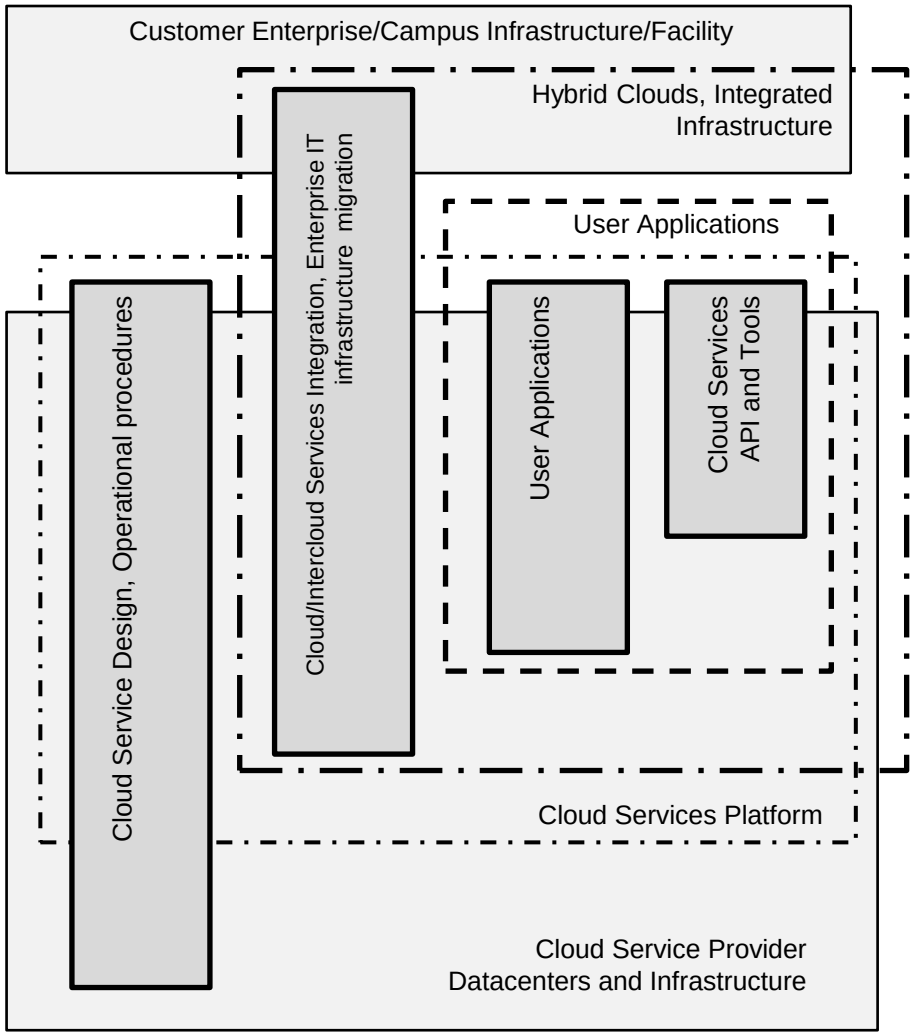
Data Links

Relations Course Components and CSM



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	



Proposed Cloud Computing Course Structure

Basic parts & Advanced parts

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 overview

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

Part 3.1. Major cloud provider platforms: Amazon AWS, Microsoft Azure, GoogleApps, etc

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

Part 4. Cloud middleware platforms: Architecture, platforms (OpenStack, OpenNebula), 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

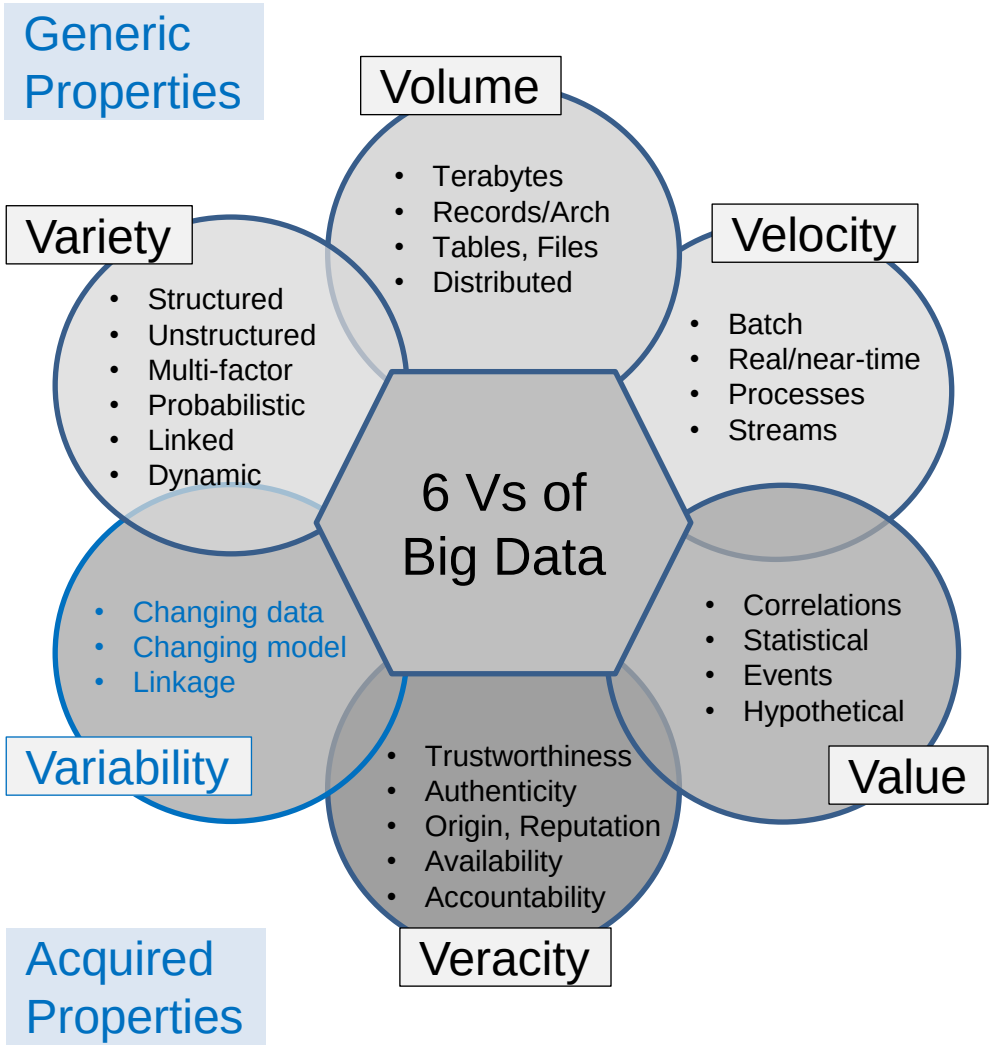
Part 7.2. Security services design in clouds; security models and Identity management

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

Dig Data Architecture Framework (BDAF)

- As a basis for Education and Training in Big Data or Data Intensive Science

Big Data Properties and Definition



5 parts Big Data definition

(1) Big Data Properties: 6V

(2) New Data Models

(3) New Analytics

(4) New Infrastructure and Tools

(5) Source and Target

Big Data Definition: From 5+1V to 5 Parts (1)

(1) Big Data Properties: 5V
- Volume, Variety, Velocity, Value, Veracity
- Additionally: Data Dynamicity (Variability)

(2) New Data Models
- Data Lifecycle and Variability
- Data linking, provenance and referral integrity

(3) New Analytics
- Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools
- High performance Computing, Storage, Network
- Heterogeneous multi-provider services integration
- New Data Centric (multi-stakeholder) service models
- New Data Centric security models for trusted infrastructure and data processing and storage

(5) Source and Target
- High velocity/speed data capture from variety of sensors and data sources
- Data delivery to different visualisation and actionable systems and consumers
- Full digitised input and output, (ubiquitous) sensor networks, full digital control

Big Data Definition: From 5V to 5 Parts (2)

Refining Gartner definition

- Big Data (Data Intensive) Technologies are targeting to process (1) high-volume, high-velocity, high-variety data (sets/assets) to extract intended data value and ensure high-veracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control; all of those demand (should be supported by) new data models (supporting all data states and stages during the whole data lifecycle) and new infrastructure services and tools that allows also obtaining (and processing data) from a variety of sources (including sensor networks) and delivering data in a variety of forms to different data and information consumers and devices.

(1) Big Data Properties: 5V
(2) New Data Models
(3) New Analytics
(4) New Infrastructure and Tools
(5) Source and Target

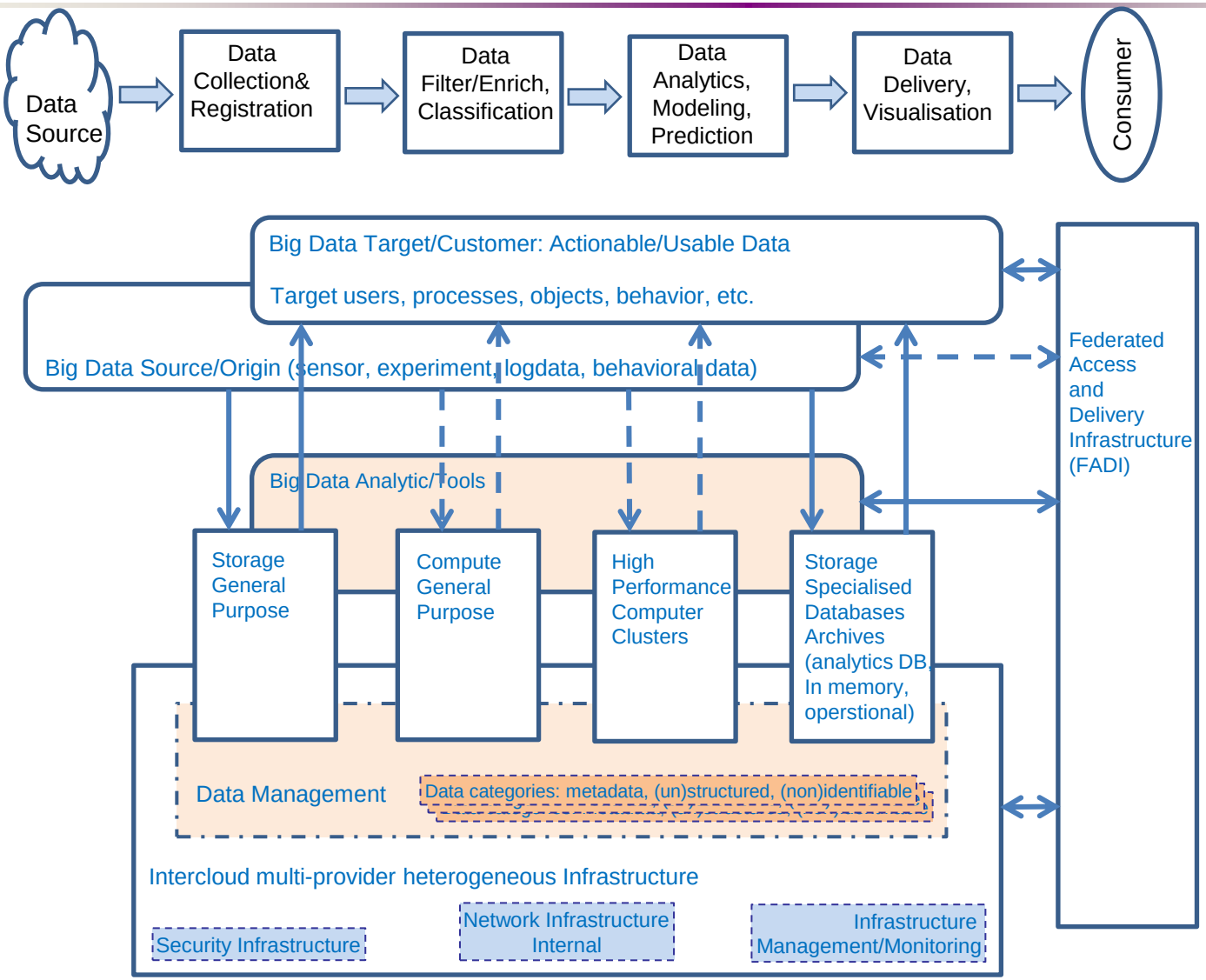
Defining Big Data Architecture Framework

- Existing attempts don't converge to consistent view: ODCA, TMF, NIST
  - See http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf
- Big Data Architecture Framework (BDAF) by UvA
  Architecture Framework and Components for the Big Data Ecosystem.
  Draft Version 0.2
  http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf
- Architecture vs Ecosystem
  - Big Data undergo a number of transformations during their lifecycle
  - Big Data fuel the whole transformation chain
    - Data sources and data consumers, target data usage
  - Multi-dimensional relations between
    - Data models and data driven processes
    - Infrastructure components and data centric services
- Architecture vs Architecture Framework (Stack)
  - Separates concerns and factors
    - Control and Management functions, orthogonal factors
  - Architecture Framework components are inter-related

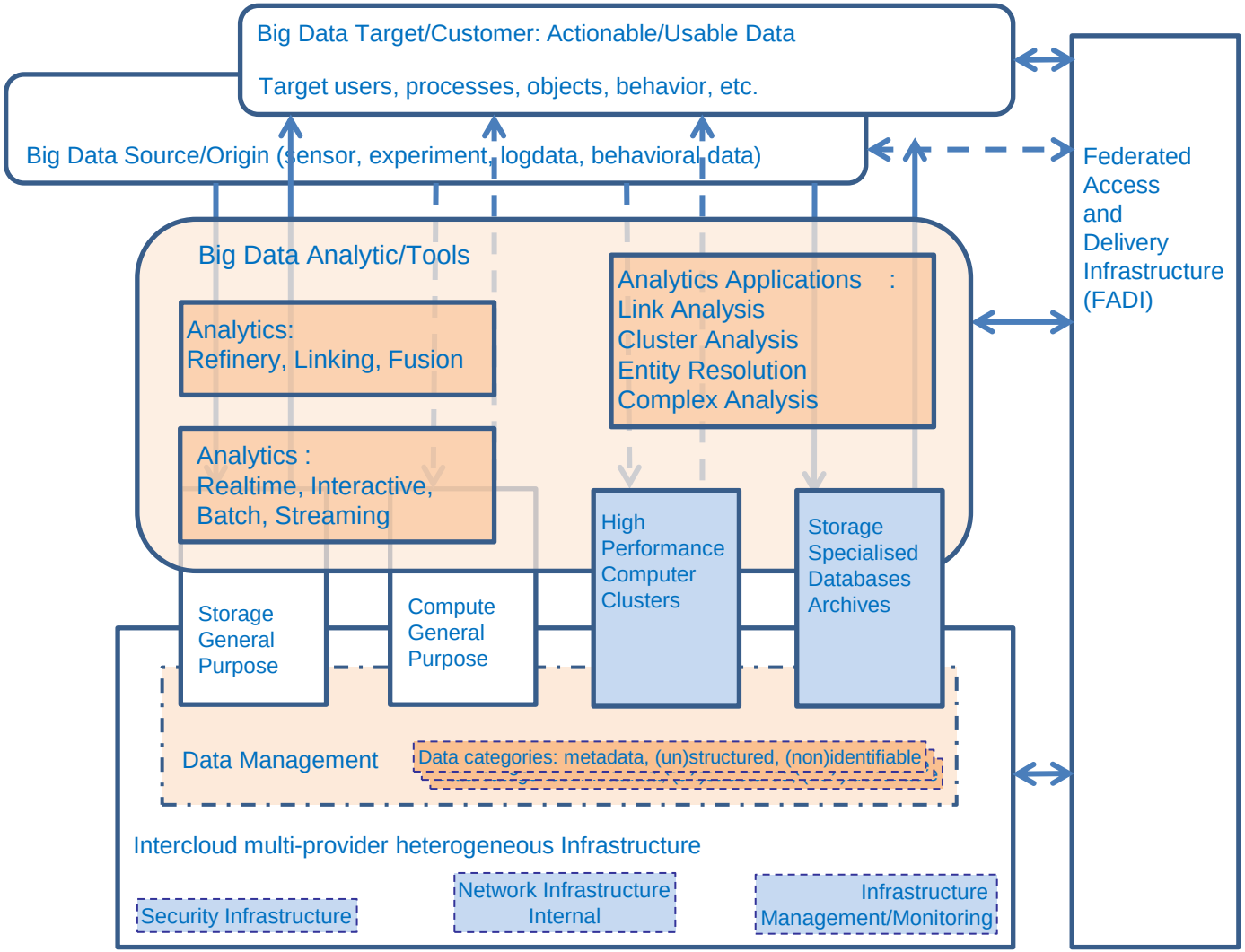
Big Data Architecture Framework (BDAF) for Big Data Ecosystem (BDE)

- (1) Data Models, Structures, Types
  - Data formats, non/relational, file systems, etc.
- (2) Big Data Management
  - Big Data Lifecycle (Management) Model
    - Big Data transformation/staging
  - Provenance, Curation, Archiving
- (3) Big Data Analytics and Tools
  - Big Data Applications
    - Target use, presentation, visualisation
- (4) Big Data Infrastructure (BDI)
  - Storage, Compute, (High Performance Computing,) Network
  - Sensor network, target/actionable devices
  - Big Data Operational support
- (5) Big Data Security
  - Data security in-rest, in-move, trusted processing environments

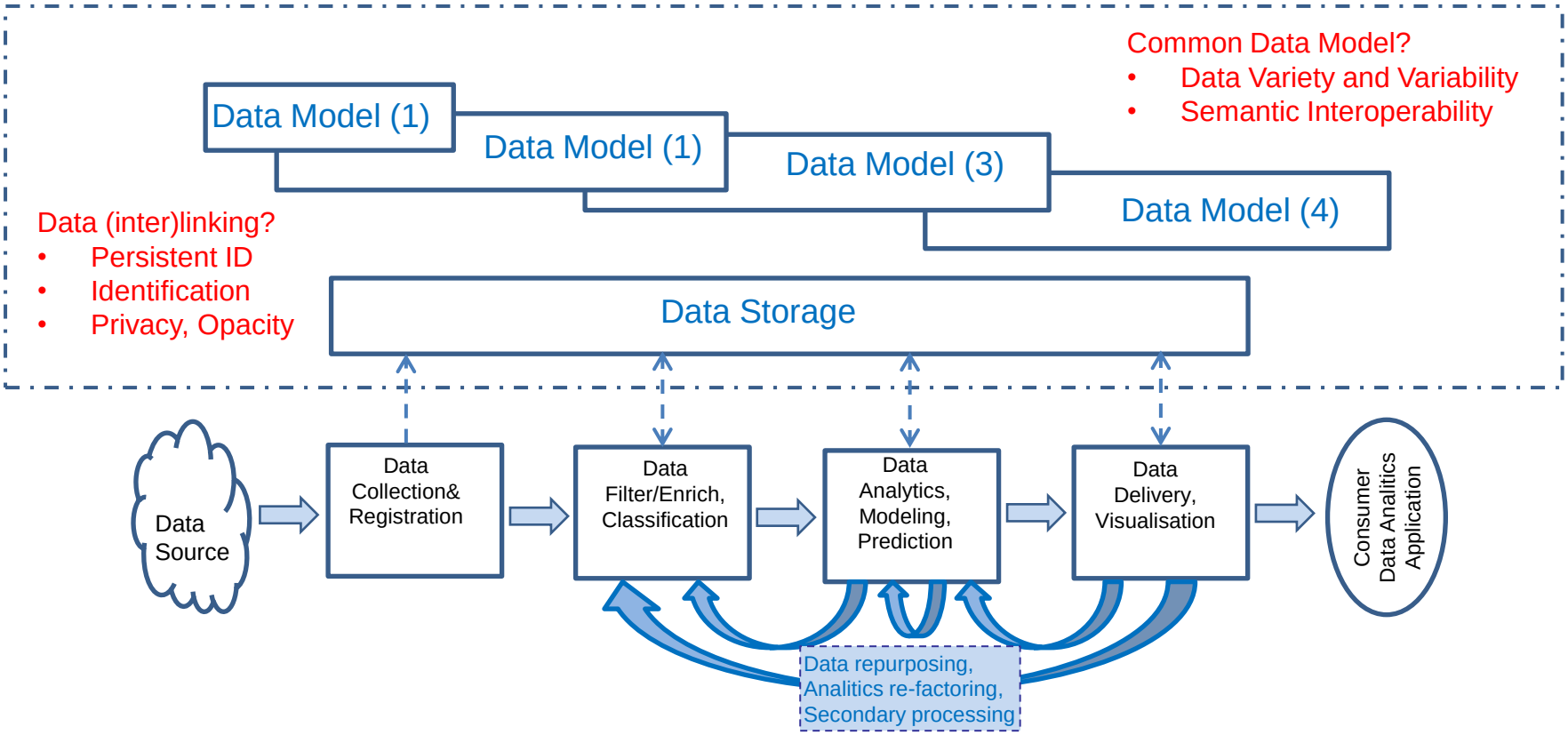
Big Data Ecosystem: Data, Lifecycle, Infrastructure



Big Data Infrastructure and Analytic Tools

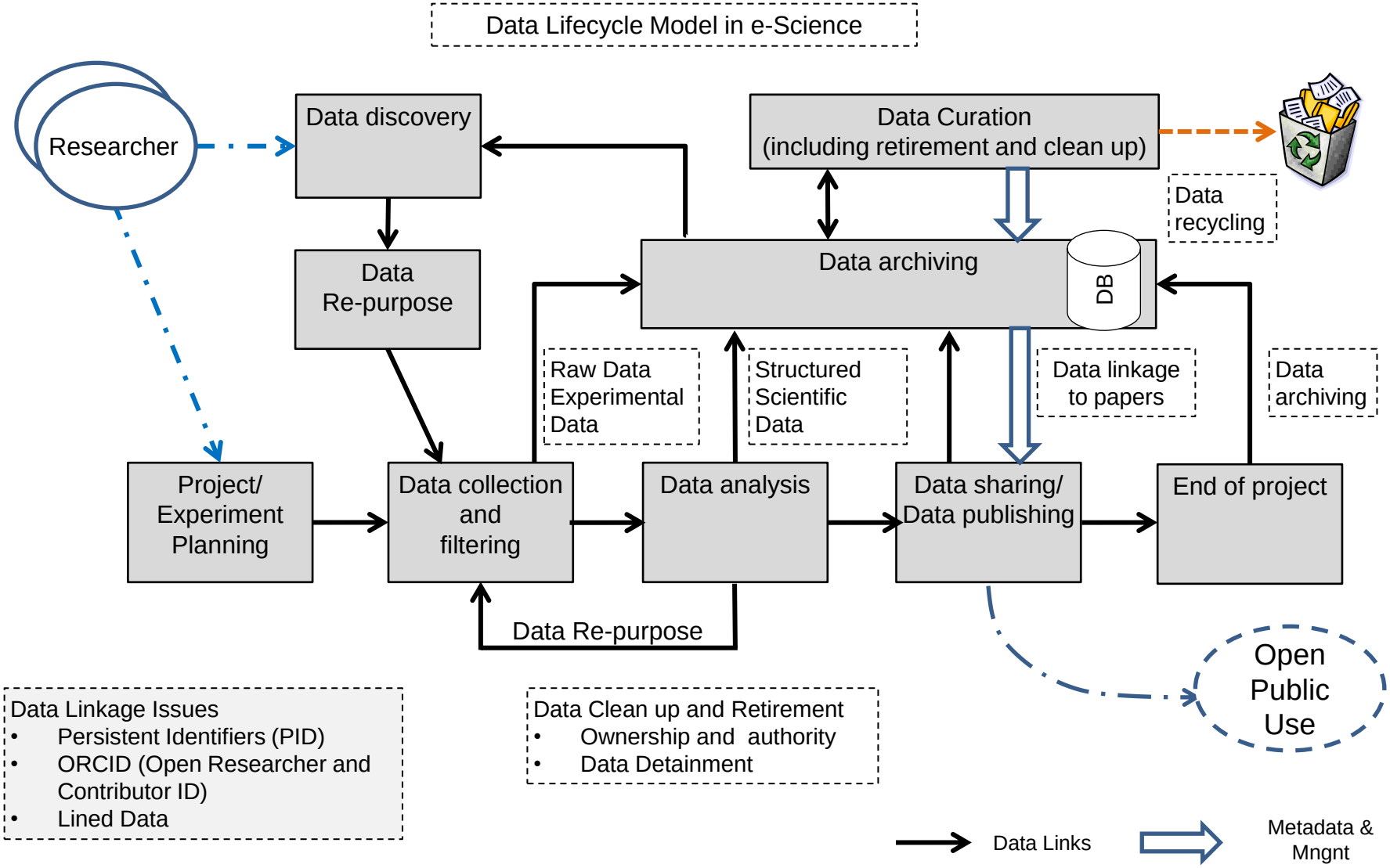


Data Transformation/Lifecycle Model



- Does Data Model changes along lifecycle or data evolution?
- Identifying and linking data
  - Persistent identifier
  - Traceability vs Opacity
  - Referral integrity

Scientific Data Lifecycle Management (SDLM) Model



BoF on Education and Skills Development in Data Intensive Science (17 Sept 2013)

- Attended by 16 representatives from universities, libraries, e-Science, data centers, research coordination bodies
- Agenda included (60 min)
  - Round of introduction and interests expression
  - Developments since the first BoF in Gothenburg
  - Presentations
    - Demystifying Data Science (Natasha Balac, SDSC)
    - Big Data in the Cloud: Research and Education (Geoffrey Fox, Indiana University Bloomington)
- Discussion on further steps
  - Priority topics to address and Interest Group establishment

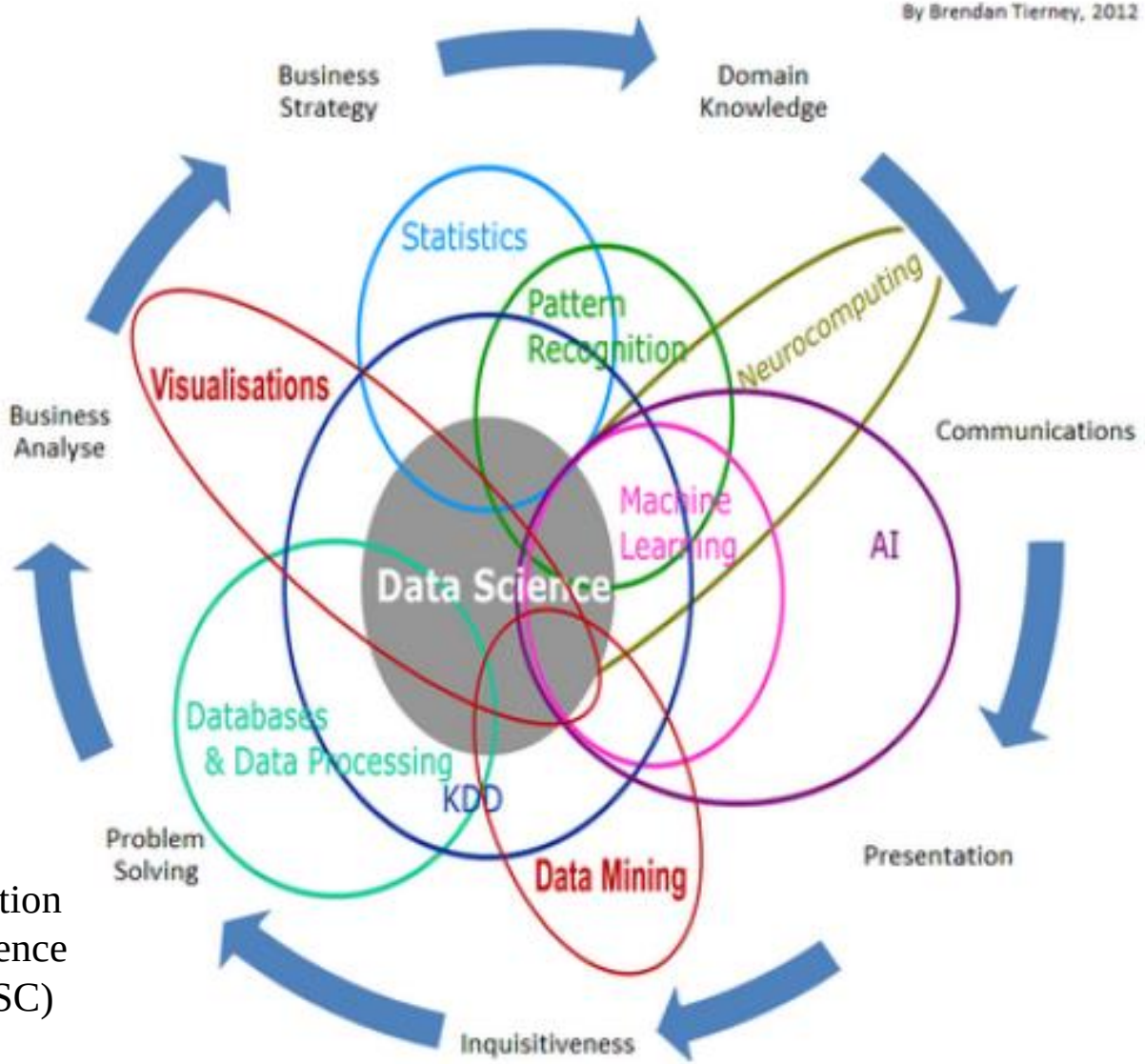
Topics discussed

- What experience do we have on component technologies to support 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

- Proceed with the formal establishment of Interest Group (IG) on Education and Skills in Data Intensive Science
  - 2 co-chairs from Europe and US volunteered – Data Analytics and e-Infrastructure
  - Seeking another co-chair from librarian or data archives community and/or AP region
  - IG scope to reflect interest of the major stakeholders: university, research, libraries, data archives, industry, subject domains – yet to identify
- Involve associations concerned with DIS/BD Education and training
  - LERU, LIBER and similar organisations in US/worldwide
  - Involve/liaise with industry – via standardisation bodies or direct contacts with leaders
- Hold the next meeting as a proposed IG
  - With more focused discussion on the community needs in defining basic skills and required knowledge for Data Science and Data Scientist
  - Reach wider and targeted community and potential stakeholders and interested parties
  - Involve universities working on the DIS education programs

Data Science Is Multidisciplinary



Slide from the presentation Demystifying Data Science (by Natasha Balac, SDSC)

Discussion

- Experience of delivering Education&Training to DW audience
- Importance of the pre-requisite and introductory knowledge module
- Possibility for profiling the course
- Need for a local base, online access and services to be negotiated and tested in advance