

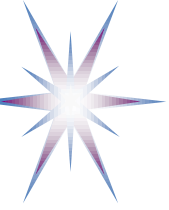
---

**BoF**

# **Education and Skills Development in Data Intensive Science**

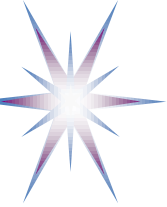
2nd RDA meeting 16-18 September 2013, Washington

(Yuri Demchenko, University of Amsterdam)



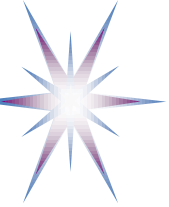
# Report to RDA2 Plenary

---



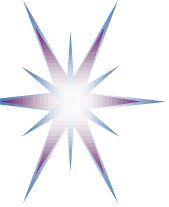
# BoF on Education and Skills Development in Data Intensive Science

- Attended by 16 representatives from universities, libraries, e-Science, data centers, research coordination bodies
- 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



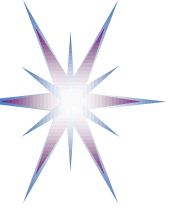
## Topics discussed

- 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

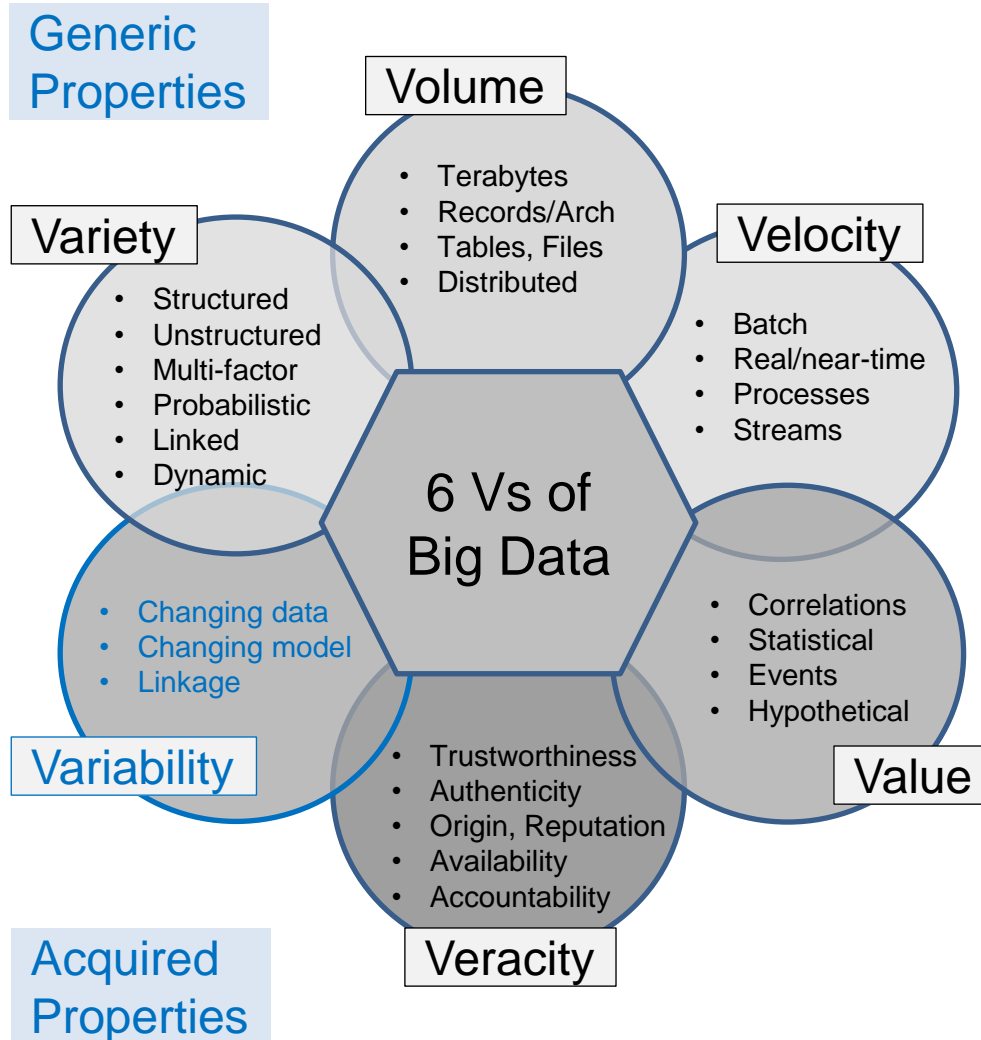


# BoF Outcome and Decisions

- 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



# Big Data Properties and Definition



## 5 parts Big Data definition

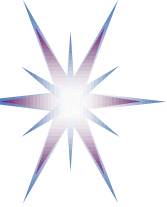
(1) Big Data Properties: 6V

(2) New Data Models

(3) New Analytics

(4) New Infrastructure and Tools

(5) Source and Target

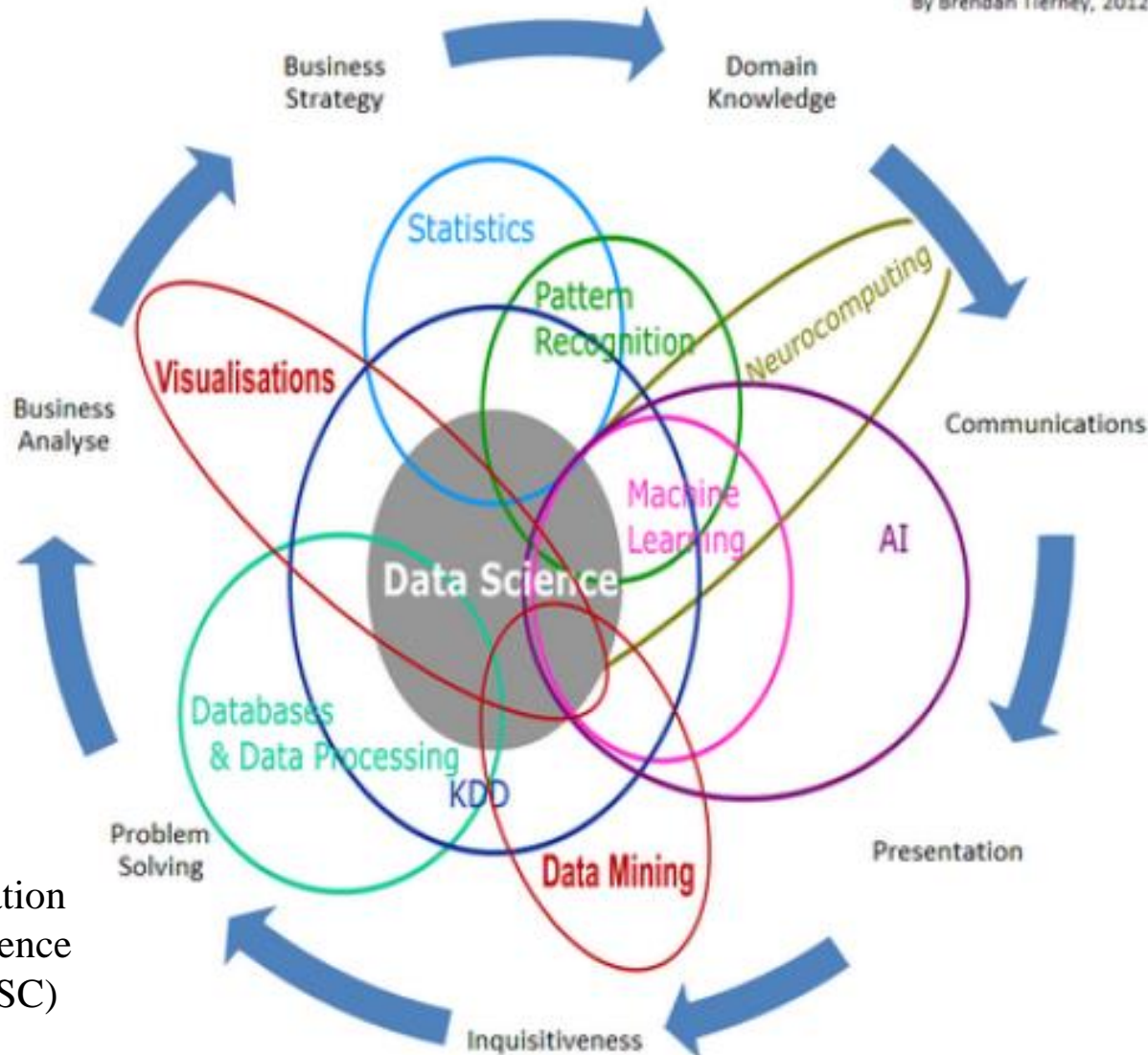


# Big Data Architecture Framework (BDAF) for Big Data Ecosystem (BDE)

- (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

# Data Science Is Multidisciplinary

By Brendan Tierney, 2012



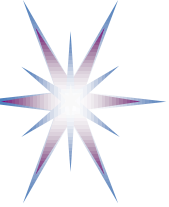
Slide from the presentation  
Demystifying Data Science  
(by Natasha Balac, SDSC)





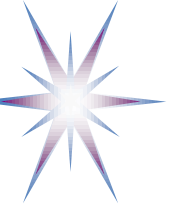
# BoF Agenda and Introductory Presentation

---



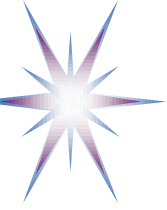
# Agenda

- Agenda bashing
- (Round of introduction)
- Overview developments since 1st BoF in Gothenburg and Introduction to discussion (Yuri Demchenko, UvA)
  - NIST Big Data WG and Big Data Architecture definition by UvA
- Presentations
  - Demystifying Big Data Science (Natasha Balac, SDSC)
  - Big Data in the Cloud: Research and Education (Geoffrey Fox, Indiana University Bloomington)
- Discussion on further steps
  - Interest Group on Education and Training in Data Intensive Science and Technologies – Drafting charter
  - Plans for 3rd RDA Plenary (March 2014, Dublin, Ireland)
- AOB



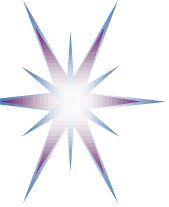
# Possible topics for discussion

- 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?
- Program development principles and example



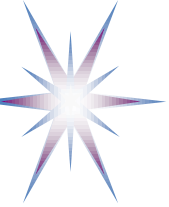
# DIST Program development principles and example

- Reuse experience from the existing related program
  - Cloud Computing Technologies and Tools
  - Theoretical Informatics, Data Analytics and Artificial Intelligence
  - Collect examples/projects and do taxonomy
- General interactive education principles
  - Common Body of Knowledge
  - Bloom's taxonomy
  - Pedagogy vs Andragogy
  - MOOC: possibilities and limitations
  - Discussion forum for online education, and Research and Reading assignments for on-campus education



# Introduction to discussion on Education and Training for Data Intensive Science

- Big Data divide and need for Professional Education and Training
- Horizon2020 Challenge 4 e-Infrastructure: Topic 11 - Skills and new professions for e-Infrastructure research data
  - Initial wording was “Skills and new professions for research data”
- Data Intensive Science and foundation technologies
  - Big Data and Data Science – Definition and Architecture Framework
- Examples Data Science / Data Intensive Science / Big Data curricula development
  - To be presented by Natasha Balac and Geoffrey Fox (by presentation)



# Horizon2020: Challenge (2.1): Development, deployment and operation of e-infrastructures

Call on Specific Challenge 2.1: Development, deployment and operation of e-infrastructures

CHALLENGE 1 – High Performance Computing (HPC)

CHALLENGE 2 - CONNECTIVITY

CHALLENGE 3 - DATA

Topic 5: Community data services

Topic 6: Managing, preserving and computing with big research data

Topic 7: e-Infrastructure for Open Access

Topic 8: Towards global data e-infrastructures - RDA

**CHALLENGE 4 – e-INFRASTRUCTURE INTEGRATION**

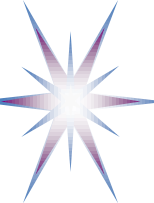
Topic 9: e-Infrastructures for virtual research environments (VRE)

Topic 10: Provisioning of core services across e-Infrastructures

**Topic 11: Skills and new professions for e-infrastructures**

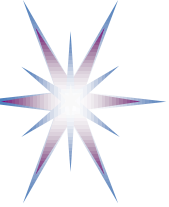
CHALLENGE 5 – POLICY AND INTERNATIONAL

Topic 12: Policy development and international cooperation



# H2020 Topic 11 - Skills and new professions for e-Infrastructure research data

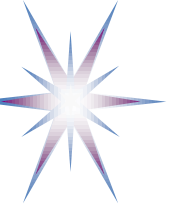
- The changing methods of (digital) science and research require that researchers, professors and students receive adequate support in
  - **computing and networking**, as well as in
  - **handling, analysing**, and
  - **storing large amounts of data and content**
- Scope of work of the emerging professions of
  - **e-infrastructure developers, integrators, operators, research technologists, data scientists and data librarians**
- Development of appropriate **curricula, training and skills**
- **Training opportunities** should be available at all levels and for all communities potentially engaged in research related activities



# H2020 Topic 11 Activities

- **Support the establishment of these professions** as distinct professions from that of a researcher and a traditional computer specialist/engineer
- **Create a reference model and a Common Body of Knowledge (CBK)** which defines their competencies, supported by
  - **case studies and best practices** relating to e-Infrastructure skills,
  - **human resources management,**
  - **support tools and related institutional practices** (community capability)
- **Defining or updating university curricula** for the e-infrastructure competences mentioned above, and promoting their adoption
- **Support networking and information sharing** among already practicing e-infrastructure experts, research technologists, data scientists and data librarians working in research institutes and in higher education

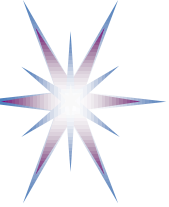




# Big Data Technology Domain

---

- Big Data Architecture Framework definition

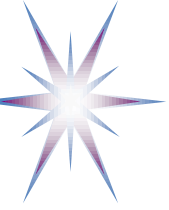


# Big Data Definition

- 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 \*)*

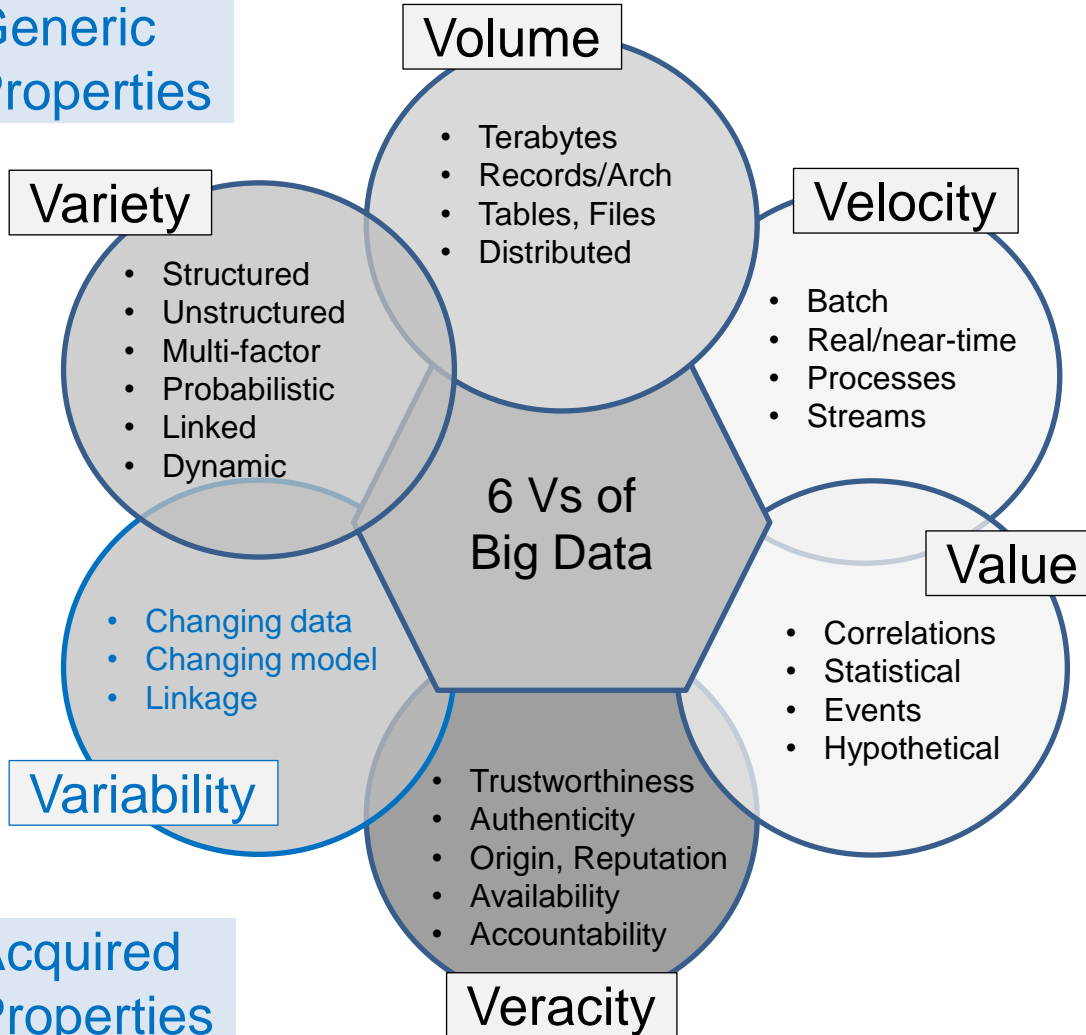
*\*) The Fourth Paradigm: Data-Intensive Scientific Discovery.  
Edited by Tony Hey, Stewart Tansley, and Kristin Tolle.  
Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4*

- 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"



# Improved: 5+1 V's of Big Data

## Generic Properties



## Acquired Properties

## Generic Big Data Properties

- Volume
- Variety
- Velocity

## Acquired Properties (after entering system)

- Value
- Veracity
- Variability



# 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.

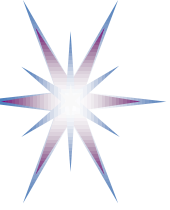
(1) Big Data Properties: 5V

(2) New Data Models

(3) New Analytics

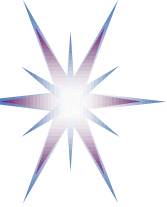
(4) New Infrastructure and Tools

(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](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

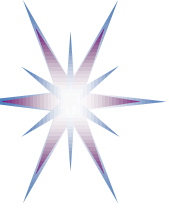
- Big Data Applications
  - Target use, presentation, visualisation

## (4) Big Data Infrastructure (BDI)

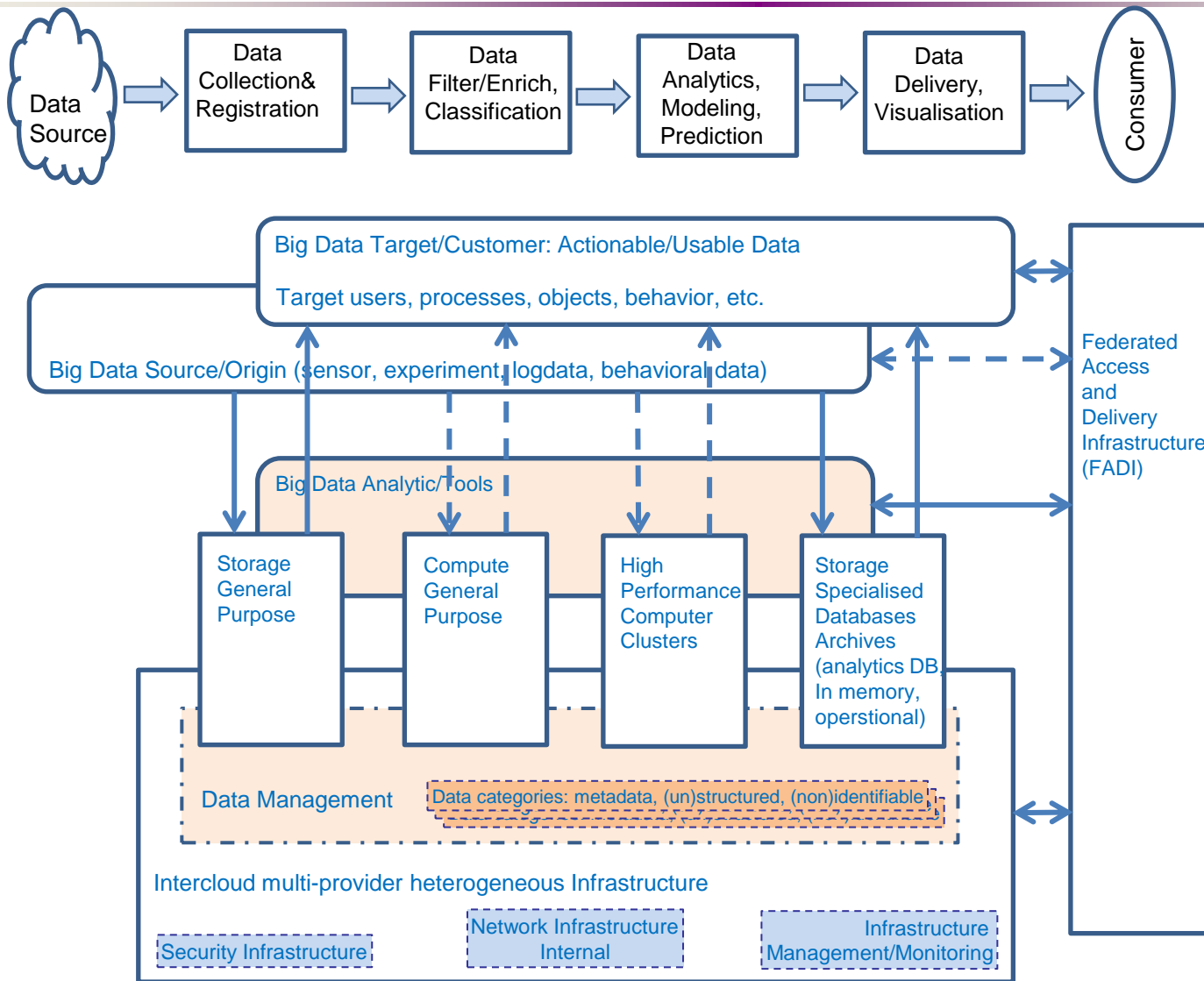
- Storage, Compute, (High Performance Computing,) Network
- Sensor network, target/actionable devices
- Big Data Operational support

## (5) Big Data Security

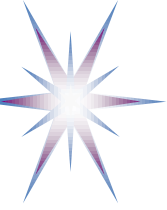
- Data security in-rest, in-move, trusted processing environments



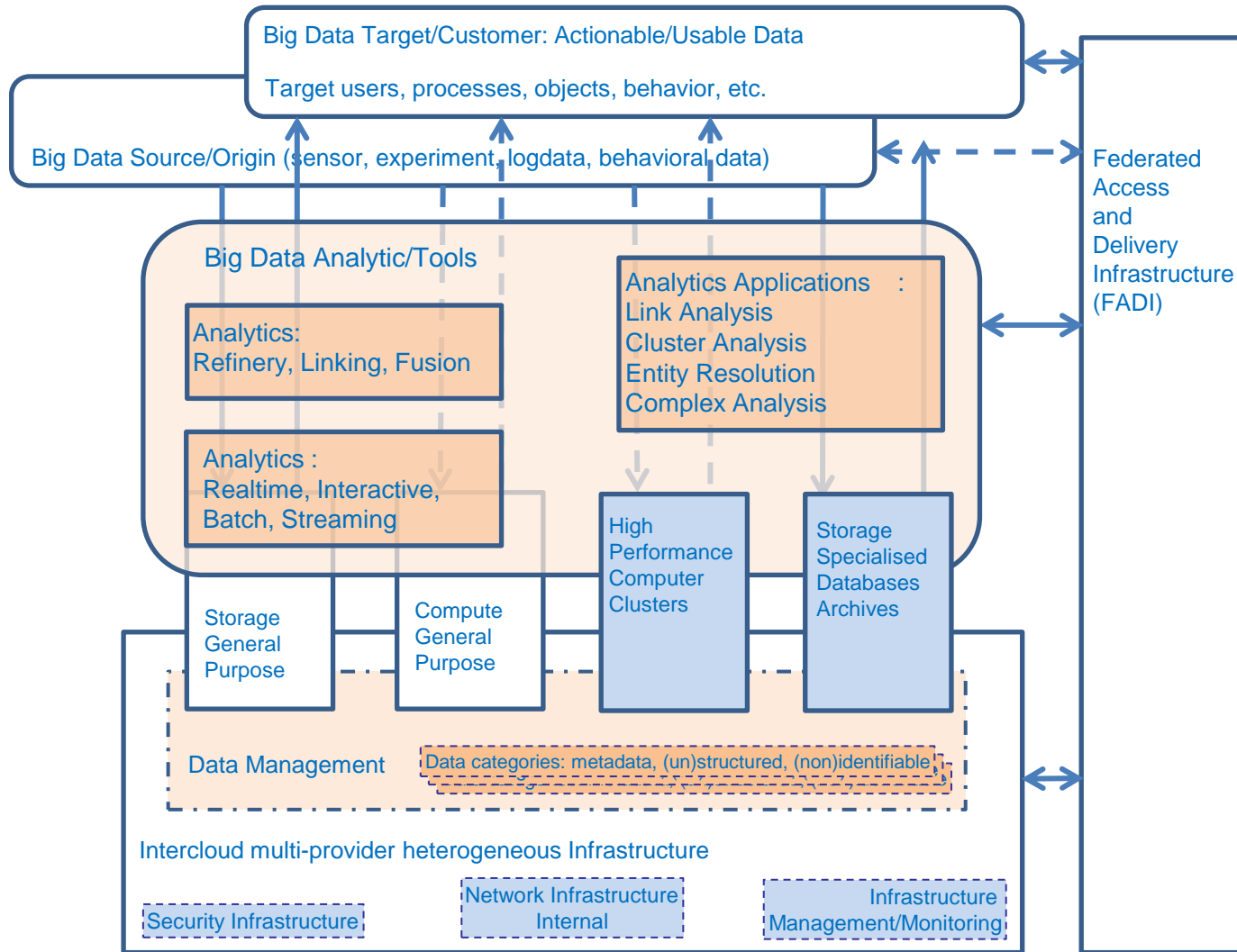
# Big Data Ecosystem: Data, Lifecycle, Infrastructure

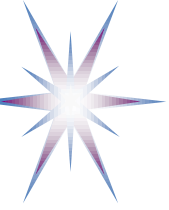




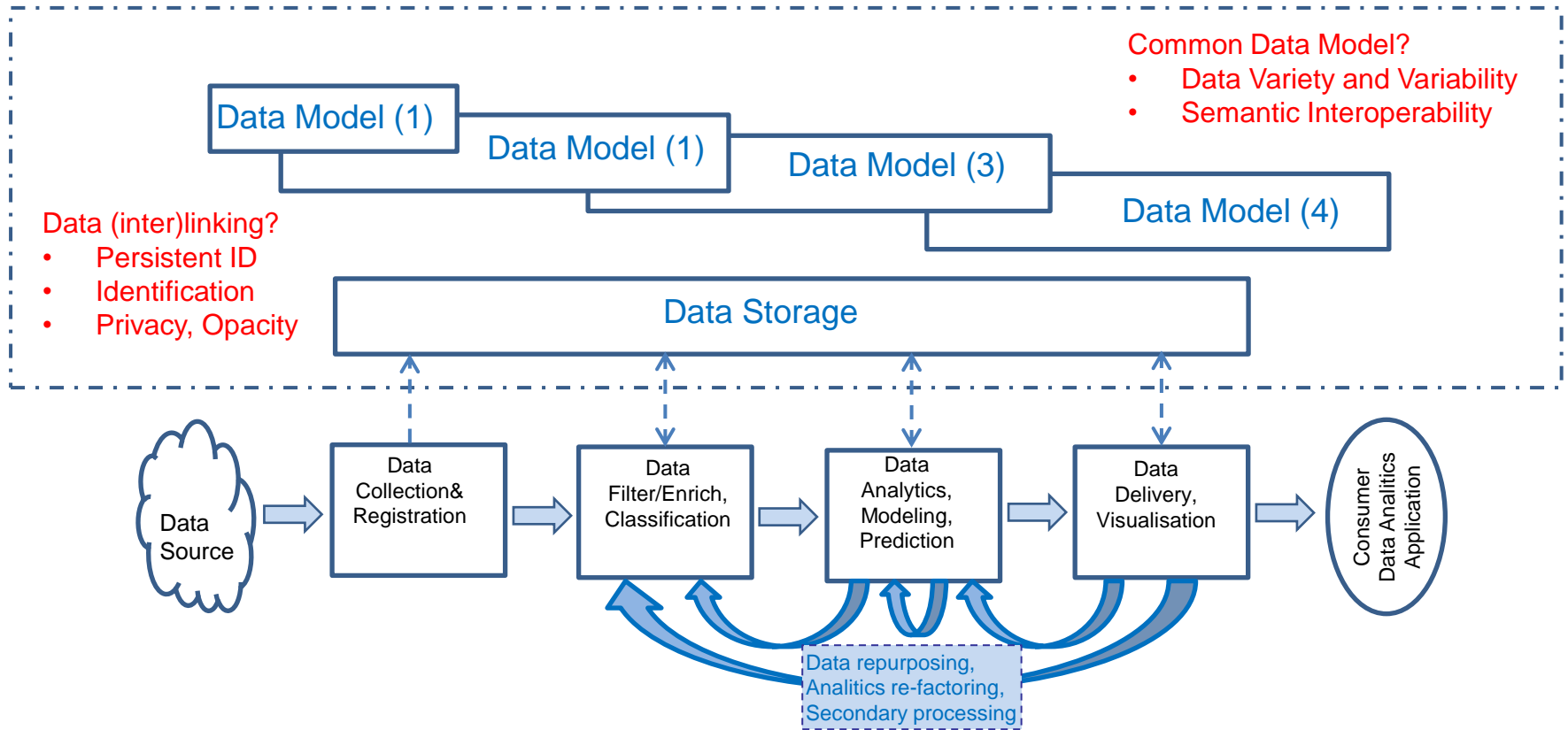


# Big Data Infrastructure and Analytic Tools



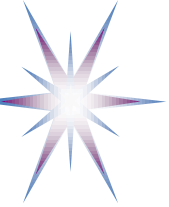


# Data Transformation/Lifecycle Model



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



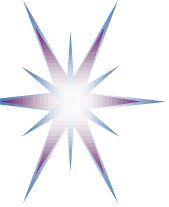


# Example: Cloud Computing Curriculum Development

Presented at the BoF during the 1st RDA meeting 18-20 March 2013 in Gothenburg

<http://www.uazone.org/demch/presentations/rda2013-göthenburg-bof-education-skills-v02.pdf>

- Cloud Computing as enabling technology for Scientific Data Infrastructure (SDI) and Big Data Infrastructure
  - Needs to be a part of Big Data technology education
- Cloud Computing Common Body of Knowledge (CBK)
- Course instructional approach: Bloom's Taxonomy and Andragogy
- Course structure Cloud Computing technologies and services design



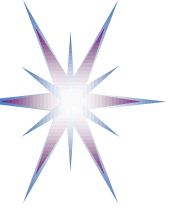
# Example: Common Body of Knowledge (CBK) in Cloud Computing

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

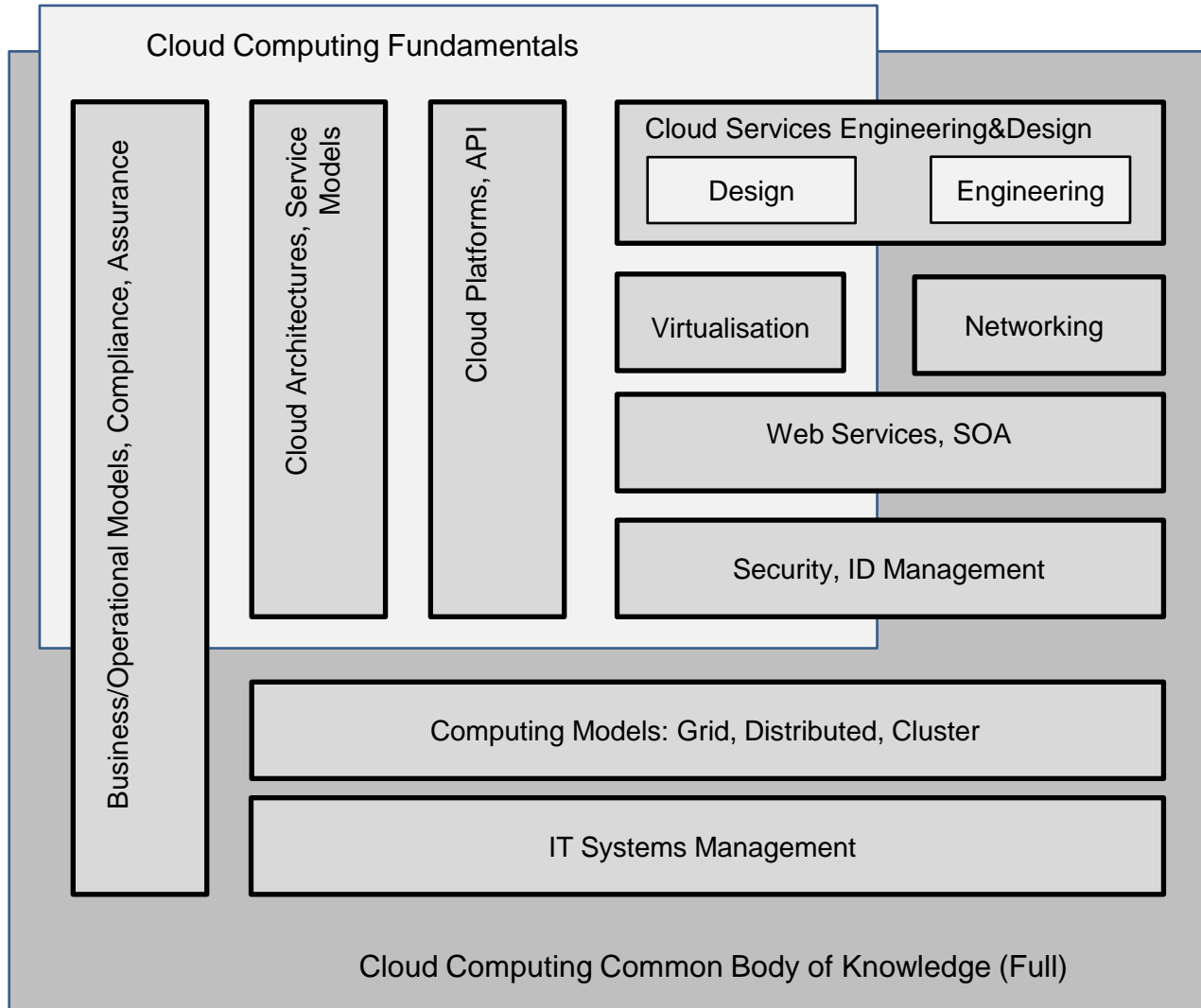
- Still in development but already piloted by some companies, including industry certification program (e.g. IBM, AWS?)

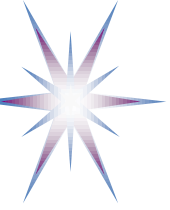
CBK Cloud Computing elements

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





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

Taxonomy  
Cognitive  
Domain [3]

Knowledge

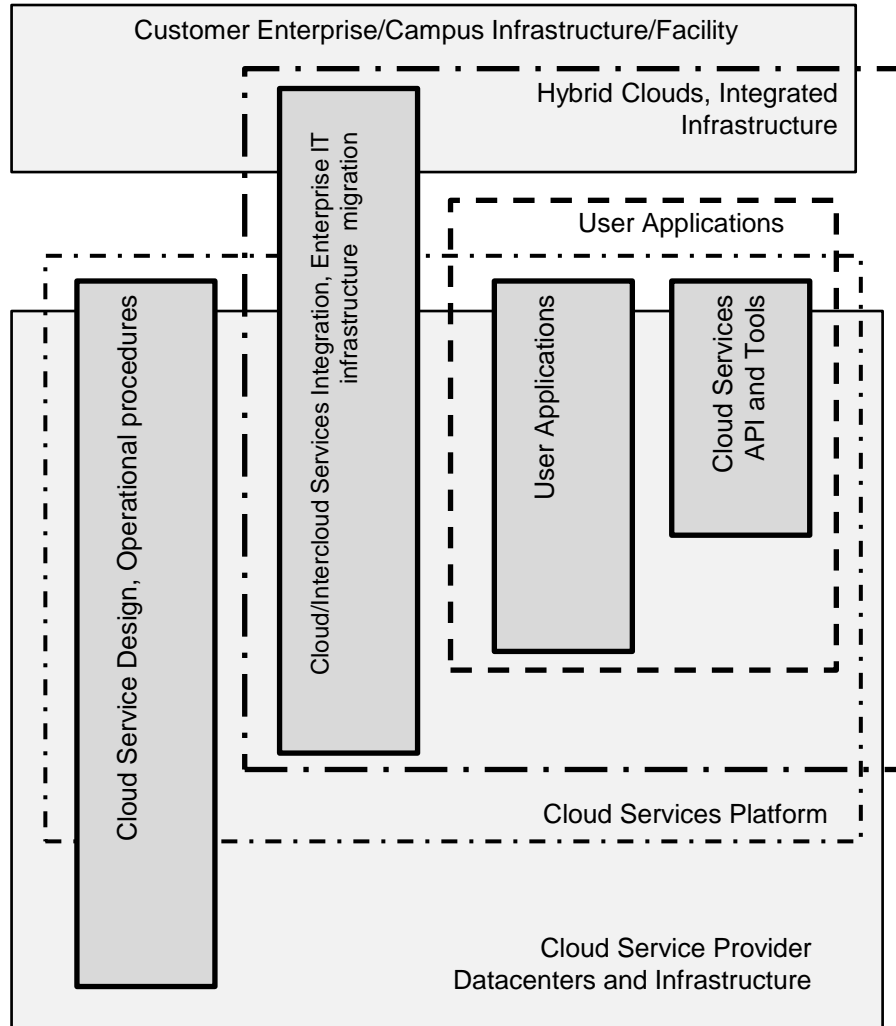
Comprehension

Application

Analysis

Synthesis

Evaluation



Taxonomy  
Professional  
Activity Domain

Perform standard tasks,  
use standard API and  
Guidelines

Create own complex  
applications using  
standard API (simple  
engineering)

Integrate different  
systems/components,  
e.g. provider and  
enterprise infrastructure

Extend existing services,  
design new services

Develop new architecture  
and models, platforms  
and infrastructures