Big Data and Education:

Education and Skills Development in Big Data and Data Intensive Science

Yuri Demchenko

SNE Group, University of Amsterdam

ISO/IEC SGBD Big Data Technologies Workshop

Part of ISO/IEC Big Data Study Group meeting

13-16 May 2014

Introduction to discussion on Education and Training for Big Data (or Data Intensive Science and Technologies)

- Horizon2020: Education and skills development for Research e-Infrastructure

- Data Science education programs development by universities
  - Big Data and Data Science – Not only Data Mining and Data Analytics
  - Examples Data Science / Data Intensive Science / Big Data curricula development
- Initiatives and developments in US and EU
  - HPC University in US

- Initiatives and developments at the University of Amsterdam (UvA)
  - Data Science Master program development
  - **EDISON Initiative to coordinate Data Science curriculum development**

Topics for Discussion

- Big Data divide and need for Professional Education and Training
  - Between scientific domains
  - Between IT and non-IT
  - Between countries

- How to share experience between universities that have already started development of Data Science programs?
  - What experience do we have on component technologies?
  - Existing instructional and educational concepts and technologies
    - Top down approach vs bottom up approach
    - Project based collaborative education

- Education and Training program development approaches
  - Targeting different communities and research domains
  - Technical vs non-technical vs subject domains

Horizon2020: Education and skills development for Research e-Infrastructure (1)

WP2014-2015 European research infrastructures (including e-Infrastructures)

- EINFRA-4-2014 – Pan-European High Performance Computing infrastructure and services
- EINFRA-5-2015 – Centres of Excellence for computing applications
  - … research in HPC applications; and addressing the skills gap in computational science
- INFRASUPP-3-2014 – Strengthening the human capital of research infrastructures
  - The skills and expertise specifically needed to construct, operate and use research infrastructures successfully therefore are not widely available
- INFRASUPP-4-2015 – New professions and skills for e-infrastructures
  - See next slide

Horizon2020: Education and skills development for Research e-Infrastructure (2)

INFRASUPP-4-2015 – New professions and skills for e-infrastructures

(1) Defining or updating **university curricula** for the e-infrastructure competences mentioned above, and promoting their adoption.

(2) Developing and executing **training programmes** (including for lifelong learning) for the above mentioned professionals working as part of a team of researchers or supporting research teams.

(3) Support the establishment of these professions as distinct professions from that of a researcher. Create a **reference model which defines their competencies**, supported by case studies and best practices relating to e-infrastructures skills, human resources management, support tools and related institutional practices. Develop alternatives means for recognising non-research contributions by research technologists and data scientists.

(4) Support **networking and information sharing** among already practicing e-infrastructure experts, research technologists, computation experts, data scientists and data librarians working in research institutes and in higher education.

(5) Awareness raising activities; establish and promote e-infrastructures community champions to advocate on new jobs and skills needs at schools, universities and scientific communities.

Data Science education programs development by universities

- Data Science programs in US and Europe – Info page (Updated 5 Feb 2014) http://whatsthebigdata.com/2012/08/09/graduate-programs-in-big-data-and-data-science/
  - Indiana University (Prof. Geoffrey Fox)
  - San Diego Supercomputing Center (SDSC)
  - HPC University in US+
  - Many others
- Currently most of programs are refactored/derived from Data Mining and Analytics, Machine Learning, Business Analytics
  - Big Data and Data Science – Not only Data Mining and Data Analytics
  - Big Data require e-Infrastructure to support data processing, storing and management
- Examples Data Science / Data Intensive Science / Big Data curricula development
  - University of Stavanger (UiS) and Purdue University
  - University of Liverpool and Laureate Online Education
  - University of Amsterdam and Vrij Univ Amsterdam
- Numerous initiatives to provide training to non-IT community
  - Librarians, Data archivists, etc.
  - Data Intelligence 4 Librarians designed by 3TU.Datacentrum and DANS http://dataintelligence.3tu.nl/en/home

HPC University (HPCU) in US

HPC University is a Virtual Organization whose membership is open to all organizations that would like to contribute to the preparation of current and future generations of HPC practitioners

http://hpcuniversity.org/

- **XSEDE**, NCSA Blue Waters, Shodor Education Foundation, Inc., **Ohio Supercomputer Center**, San Diego Supercomputer, ACM SIGHPC, Cyprus Institute, Partnership for Advanced Computing in Europe (PRACE)
- HPCU actively seeks participation from all sectors of the HPC community to:
  - assess the learning and workforce development needs and requirements of the community,
  - catalog, disseminate and promote peer-reviewed and persistent HPC resources,
  - develop new content to fill the gaps to address community needs,
  - broaden access by a larger and more diverse community via a variety of delivery methods,
  - and pursue other activities as needed to address community needs

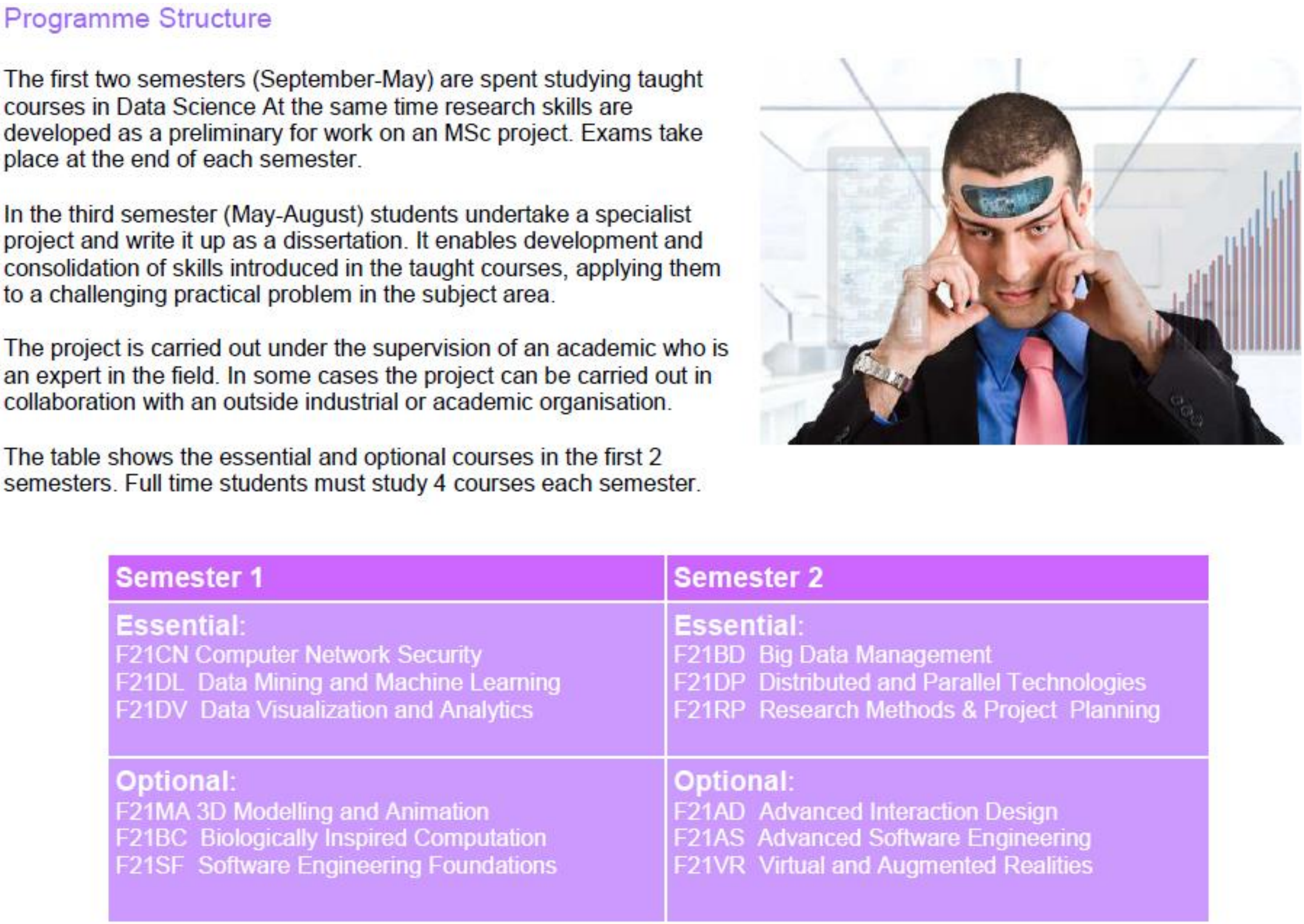
Initiatives and Developments in EU

- LERU (League of European Research Universities)
  - LERU Roadmap for Research Data
    - Prepared by the LERU Research Data Working Group
    - Valuable contribution from LIBER, UKOLN
- EGI (European Grid Initiative) Education and Training initiative
- EUDAT Training program on Research Data and tools
- EDISON Initiative (UvA, EGI, others)
  - Education for Data Intensive Science to Open New science frontiers
- Initiatives from industry and commercial companies to support Data Science and Big Data education
  - E.g. Engineering Group (HQ Italy) & Politecnico Turino, KPMG (NL) & UvA, Elsevier/LexisNexis (& UvA)

EGI Partners to cooperate on Edu@Skills

- National Centre for Scientific Research (CNRS, France)
- INFN (Italy) and University of Catania
- ENGINEERING (Italy) and Politecnico of Torino
- Alliance Permanent Archives (APA) and APARSEN Project
  - FTK (Formal Education and Curricula)
  - GLOBIT (Online Continuous Professional Education Portal)
  - DANS (Education and Training Resources)
  - INMARK (Spreading Excellence)
- University of Amsterdam (UvA)
  - Data Science Research Center
  - Academic Medical Center (AMC)

Data Science Master at School of Mathematical and Computer Sciences, Hariot Watt University



Programme Structure

The first two semesters (September-May) are spent studying taught courses in Data Science At the same time research skills are developed as a preliminary for work on an MSc project. Exams take place at the end of each semester.

In the third semester (May-August) students undertake a specialist project and write it up as a dissertation. It enables development and consolidation of skills introduced in the taught courses, applying them to a challenging practical problem in the subject area.

The project is carried out under the supervision of an academic who is an expert in the field. In some cases the project can be carried out in collaboration with an outside industrial or academic organisation.

The table shows the essential and optional courses in the first 2 semesters. Full time students must study 4 courses each semester.

Semester 1	Semester 2
Essential: F21CN Computer Network Security F21DL Data Mining and Machine Learning F21DV Data Visualization and Analytics	**Essential:** F21BD Big Data Management F21DP Distributed and Parallel Technologies F21RP Research Methods & Project Planning
Optional: F21MA 3D Modelling and Animation F21BC Biologically Inspired Computation F21SF Software Engineering Foundations	**Optional:** F21AD Advanced Interaction Design F21AS Advanced Software Engineering F21VR Virtual and Augmented Realities

Big Data Course at Laureate Online Education (University of Liverpool)

Seminar 1: Introduction. Big Data technology domain definition, Big Data Architecture Framework

Seminar 2: Big Data use cases from Science, industry and business

Seminar 3: Architecture Framework for Big Data Ecosystem, Big Data Infrastructure components

Seminar 4: Big Data analytic techniques, introduction to RapidMiner

Seminar 5: Understanding the processes behind Big Data

Seminar 6: Classification and forecasting techniques

Seminar 7: Big Data Security, Protection, and Privacy

Seminar 8: Integrating Big Data applications into enterprise IT infrastructure, data regulation compliance

Big Data Course at Laureate Online Education (University of Liverpool) - Details

Seminar 1: Introduction. Big Data technology domain definition, Big Data Architecture Framework

- **Topics:** Big Data and Data Intensive technologies definition; Big Data properties: Volume, Velocity, Variety, Variability, Value, And Veracity. Big Data ecosystem, data origin, data target; raw data and actionable data.

Seminar 2: Big Data use cases from Science, industry and business

- **Topics:** Big data use cases analysis from science and industry: LHC in HEP, LOFAR/SKA in astronomy, genomic research, Internet or web: target advertisement, recommender system (e.g. Netflix), Web Search, fraud detection

Seminar 3: Architecture Framework for Big Data Ecosystem, Big Data Infrastructure components

- **Topics:** Big Data Analytics Infrastructure and technology platforms, Enterprise Data Warehouses, MapReduce and Hadoop, new file systems and database architectures, NoSQL.

Seminar 4: Big Data analytic techniques, introduction to RapidMiner

- **Topics:** Introduction to analytics and different analytical techniques. What are Analytics/Data Mining/Machine Learning? Introduction to the RapidMiner toolkit. The use of simple pre-processing and Statistical techniques for modeling data will be described and implemented using RapidMiner.

Seminar 5: Understanding the processes behind Big Data

- **Topics:** Introduction to Rule Extraction Algorithms and Cluster Analysis. Evaluating data from text streams. Decision tree induction and Cluster Analysis techniques, use for enhancing business processes.

Seminar 6: Classification and forecasting techniques

- **Topics:** Machine Learning techniques, Neural Networks and Support Vector Machines, techniques used to analyze the performance of our analytic processes. Measurement techniques such as Receiver Operating Curves and Gains Charts.

Seminar 7: Big Data Security, Protection, and Privacy

- **Topics:** Big Data security and data protection; data centric security models. Big Data privacy issues, differential privacy and data re-identification.

Seminar 8: Integrating Big Data applications into enterprise IT infrastructure, data regulation compliance

- **Topics:** Enterprise business processes and data management issues; analysis, evaluation and optimisation; enterprise data applications engineering. Data protection and corresponding standards, regulation and legislation; data provenance.

EDISON Initiative

Data are becoming infrastructure themselves (Report “Riding the Wave”)

- This requires large infrastructure resources to collect, store, process and archive heterogeneous multi-faceted and linked data.
- Data centric/data driven infrastructure has to support different types of data, including text data, structured and unstructured data, relational and vector data, linked data.
- Data appear in various contexts: large number strings from experiments or sensors, in software code, music, films, publications, digital art, web pages, social media, public and business statistics, and also orphan data.
- We need data scientists with the knowledge and skill to work with existing and future data intensive infrastructure and tools.

EDISON: Issues to address in academic training

- Shape of the data infrastructure and its semantics
- Data discovery, visualization, processing (filtering)
- Infrastructure and capabilities for data analysis
- Uses of data and its implications
- Archiving, preservation, storage, permanency, data (de)selection, annotation
- Trust: Quality and reliability. Tracking data evolution. Curation. Safety.
- Legal, governance, environment, costs

Data scientists have to get familiar with the date universe and the peculiarities of this landscape.

They will work in research, specific scientific data domains, and in the private sector.

EDISON: Anticipated curriculum development

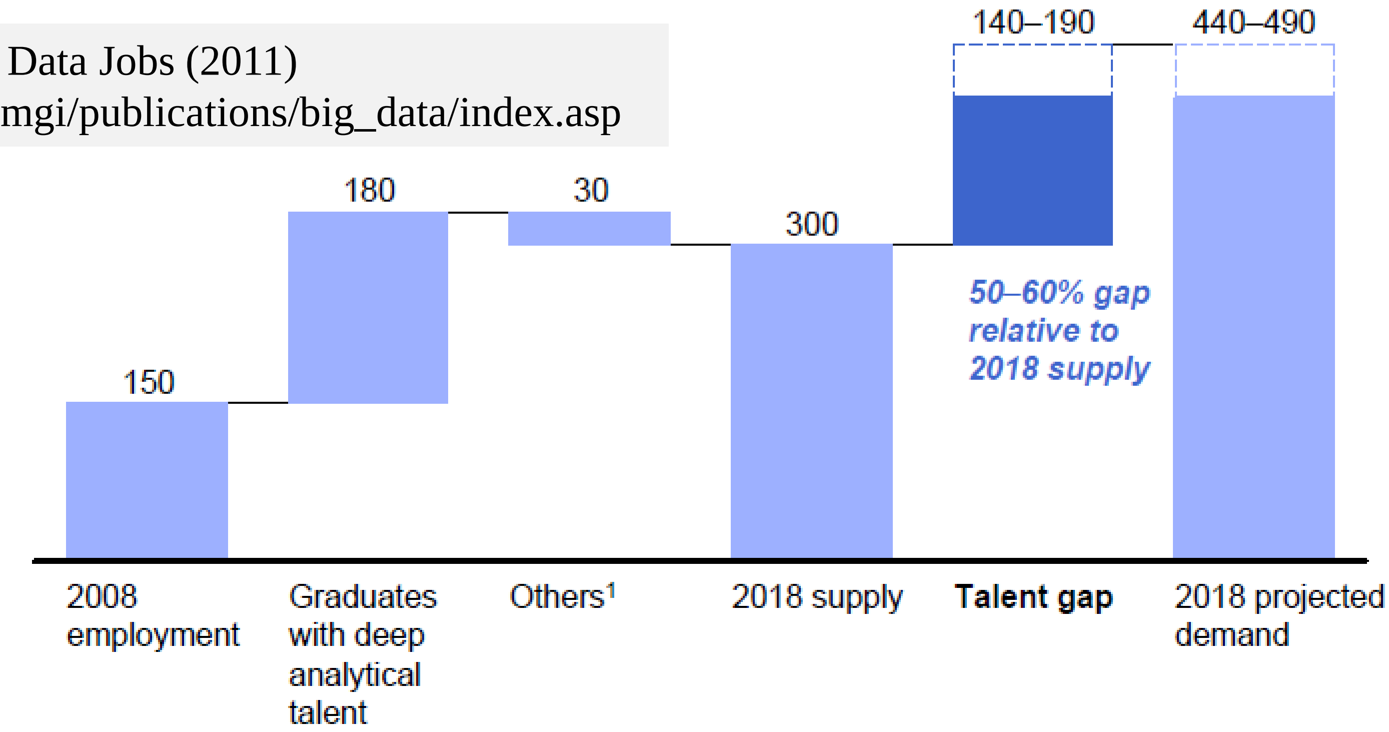
- Create a cooperative community with universities that offer bachelor and master programs in Data Science, Big Data and Data Archives.
- Consider the educational focus with a vision on the developing data universe.
- Test and adapt educational programmes.
- Address implementation strategies with respect to required expertise, methods, research training, management and financial implications.
- Promote awareness and recognition (inform the market sector; secure certification of new educational tracks).

EDISON: Next steps

- Formal establishment of the RDA Interest Group on Education and Skills Development for Research Data
- Workshop – September 2014
  - Adjacent to the next RDA4 meeting in Amsterdam
- Information exchange: Website and billboard

Data Scientist: New Profession and Opportunities

McKinsey Institute on Big Data Jobs (2011)
http://www.mckinsey.com/mgi/publications/big_data/index.asp



- There will be a shortage of talent necessary for organizations to take advantage of Big Data.
  - By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as
  - 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions

SOURCE:US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey analysis

Strata Survey Skills and Data Scientist Self-ID

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Analysing the Analysers. O'Reilly Strata Survey – Harris, Murphy & Vaisman, 2013

- Based on how data scientists think about themselves and their work
- Identified four Data Scientist clusters

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

Skills and Self-ID Top Factors

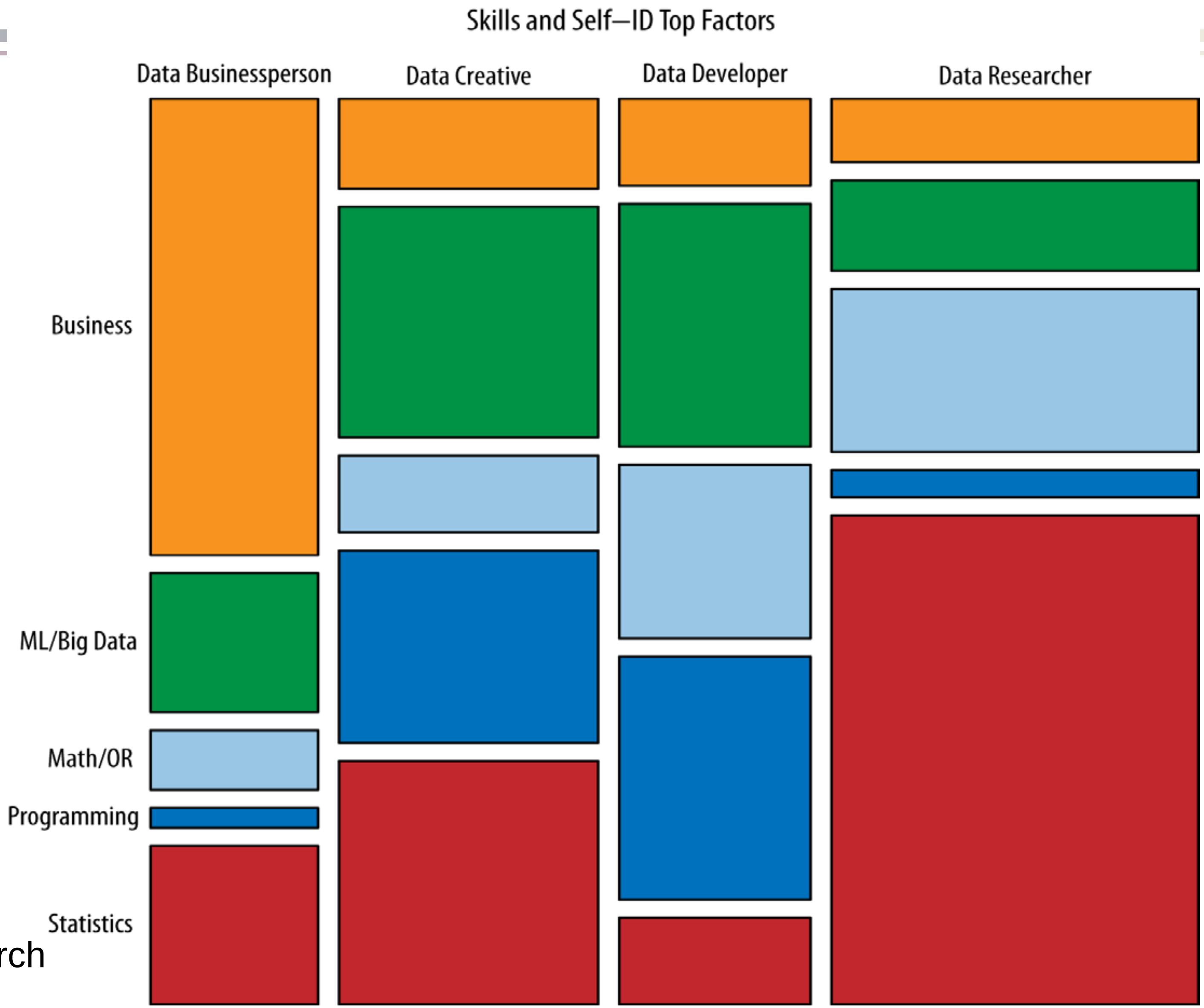
Business

ML/BigData

Math/OR

Programming

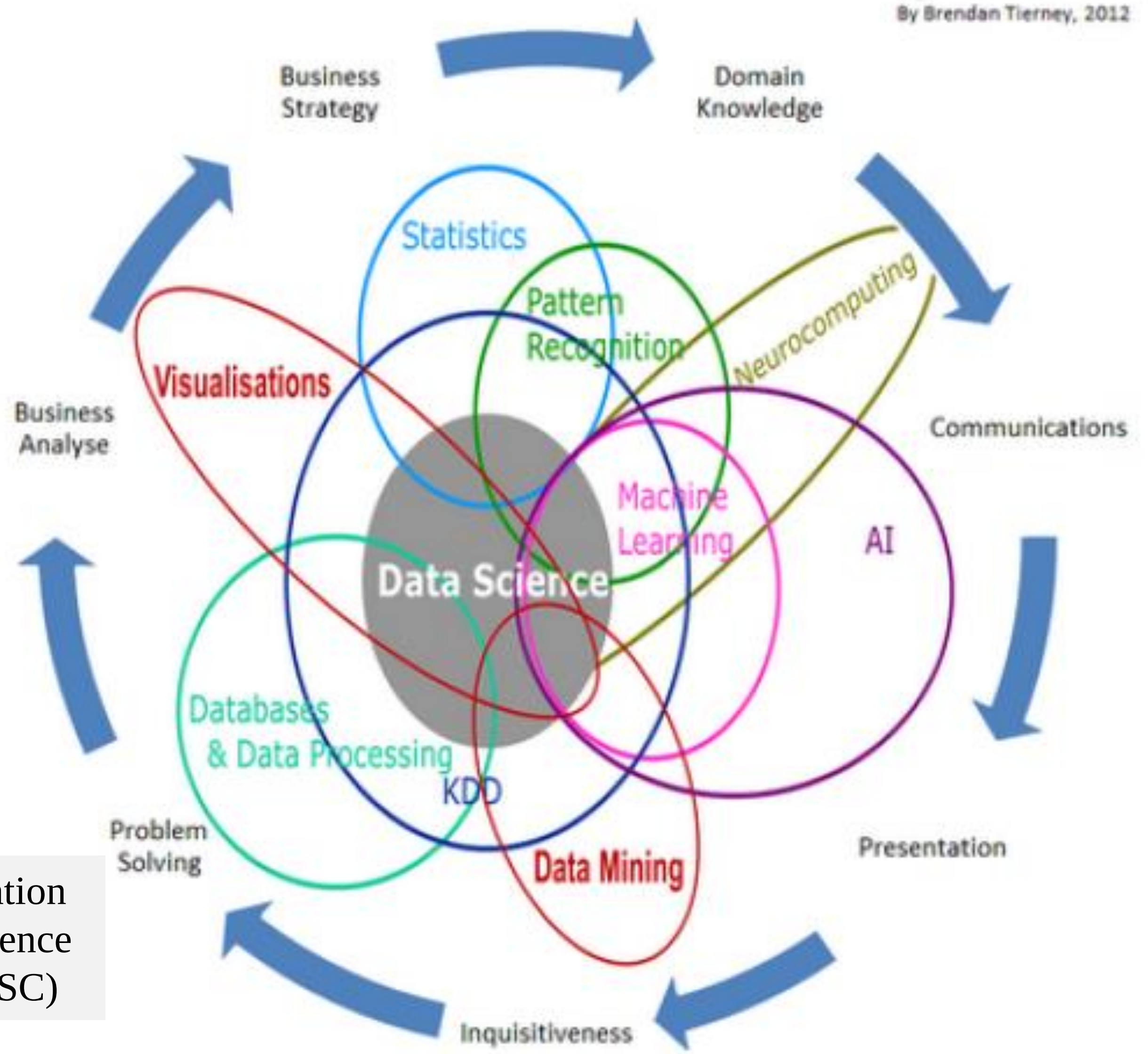
Statistics



ML – Machine Learning

OR – Operations Research

Data Science Is Multidisciplinary



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

Key to a Great Data Scientist

Technical skills (Coding, Statistics, Math)

+ Commitment +Creativity

+ Intuition

+ Presentation Skills

+ Business Savvy

= Great Data Scientist!

- How Long Does It Take For a Beginner to Become a Good Data Scientist?
  - *3-5 years according to KDnuggets survey* [278 votes total]

Questions and Discussion

Additional Information

- Example Cloud Computing course development

Common Body of Knowledge (CBK) in Cloud Computing

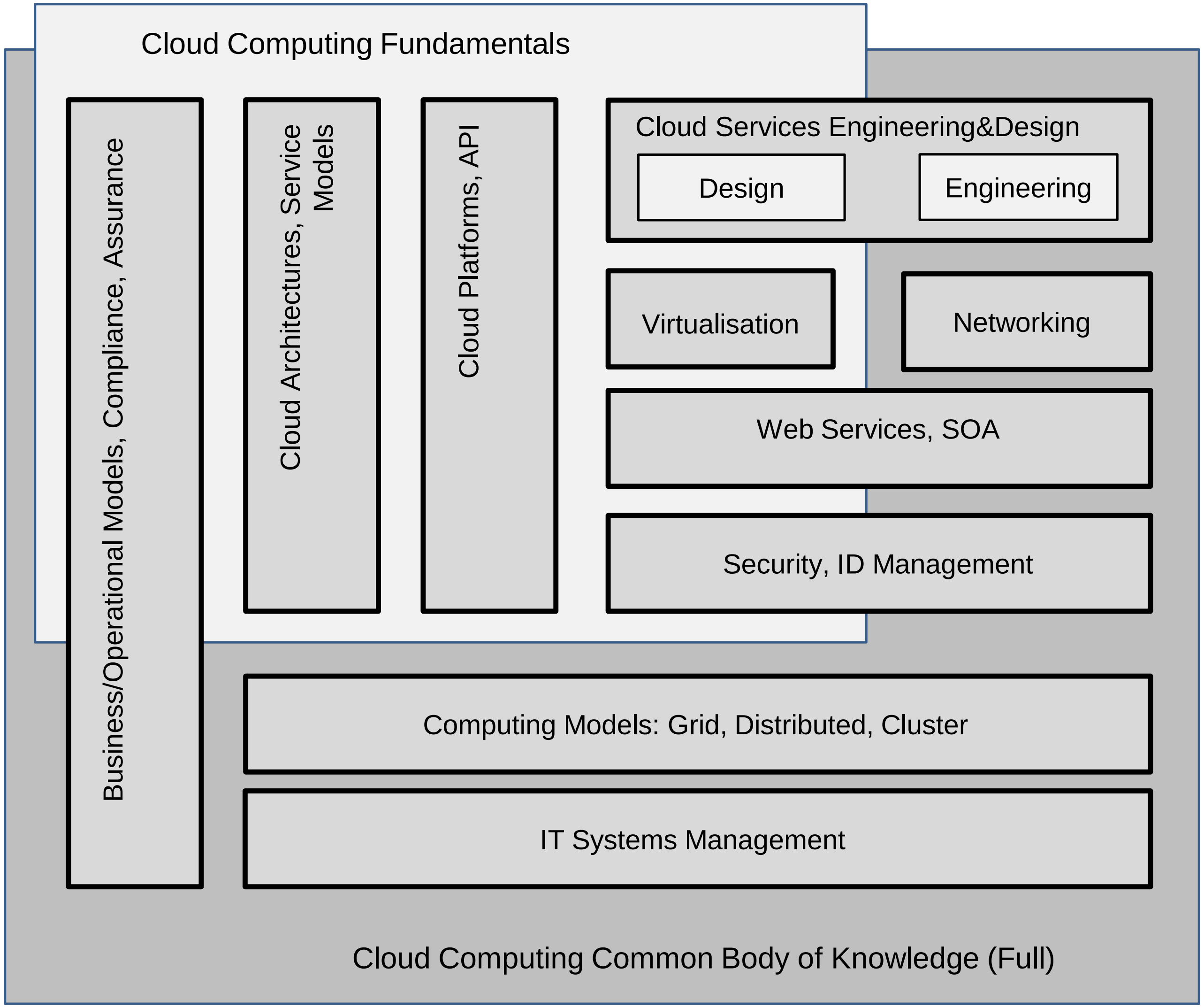
CBK refers to several domains or operational categories into which Cloud Computing theory and practices breaks down

- 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

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

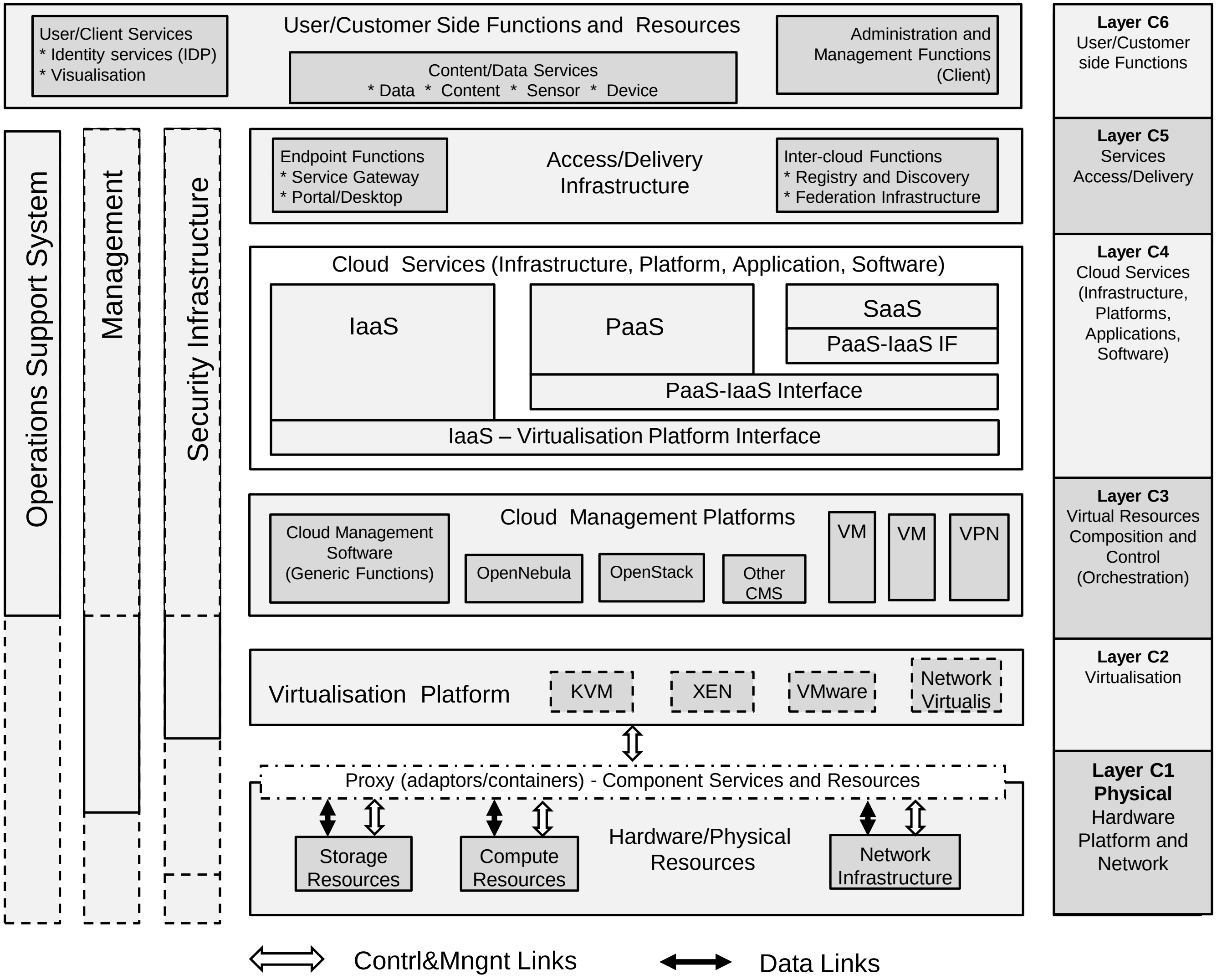
Security, ID Management

Computing Models: Grid, Distributed, Cluster

IT Systems Management

Cloud Computing Common Body of Knowledge (Full)

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, Platforms, 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

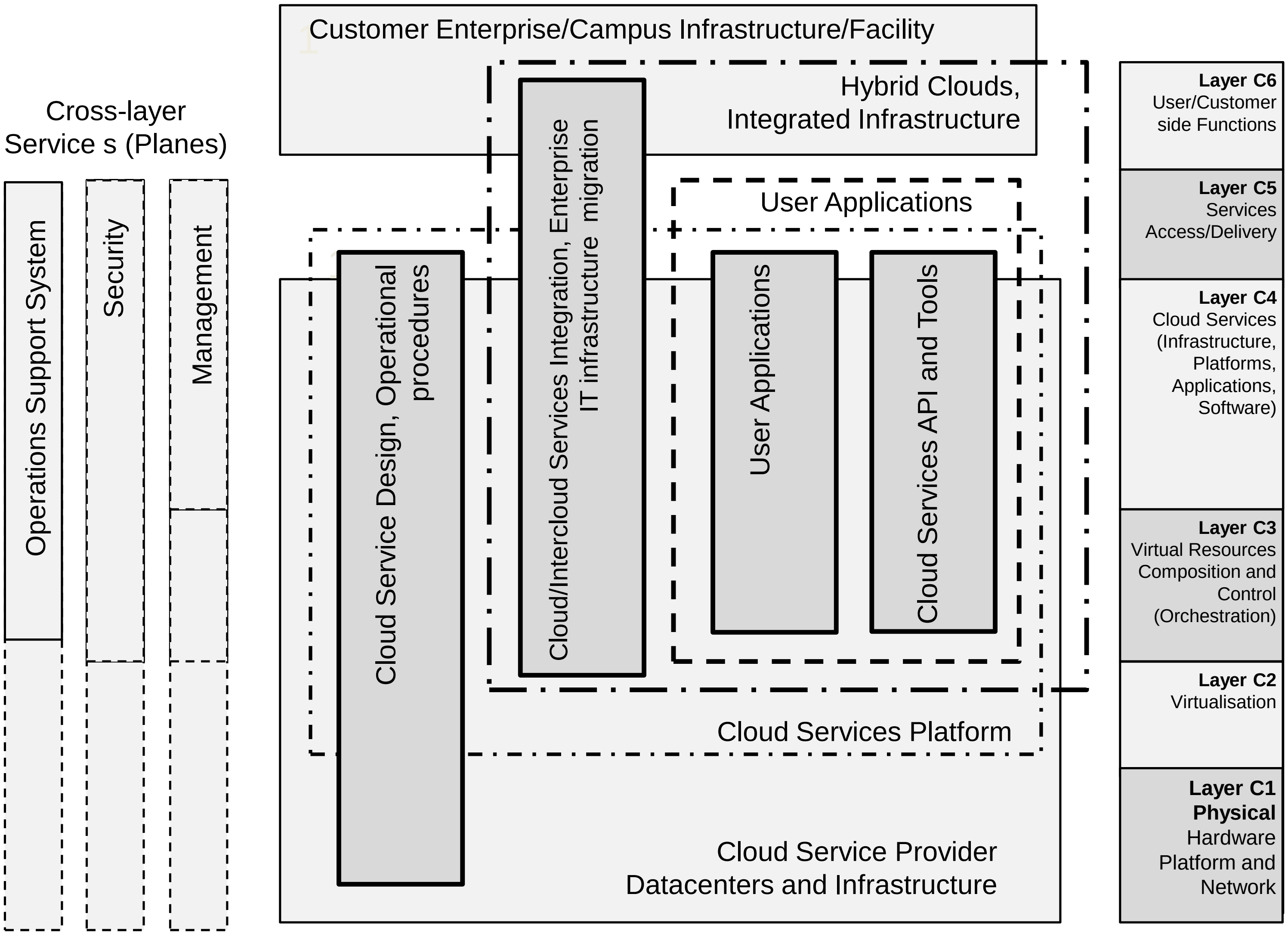
(C2) Virtualisation Layer

(C1) Hardware platform and dedicated network infrastructure

⟺ Control/ Mngnt Links

⟷ Data Links

Relations Course Components and CSM


Customer Enterprise/Campus Infrastructure/Facility
Hybrid Clouds,
Integrated Infrastructure
Cross-layer
Service s (Planes)
Operations Support System
Security
Management
User Applications
Cloud Service Design, Operational procedures
Cloud/Intercloud Services Integration, Enterprise IT infrastructure migration
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

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

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.**

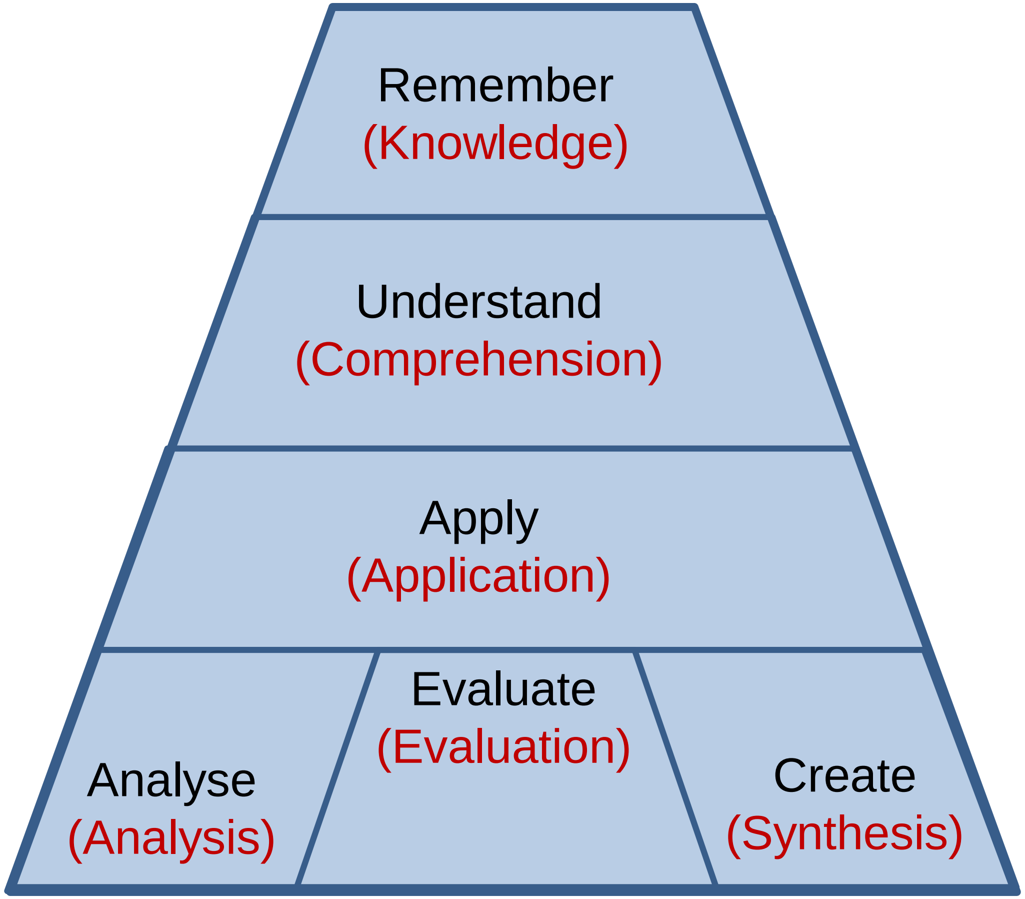
Evaluation

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?**

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



Remember (Knowledge)

Understand (Comprehension)

Apply (Application)

Analyse (Analysis)

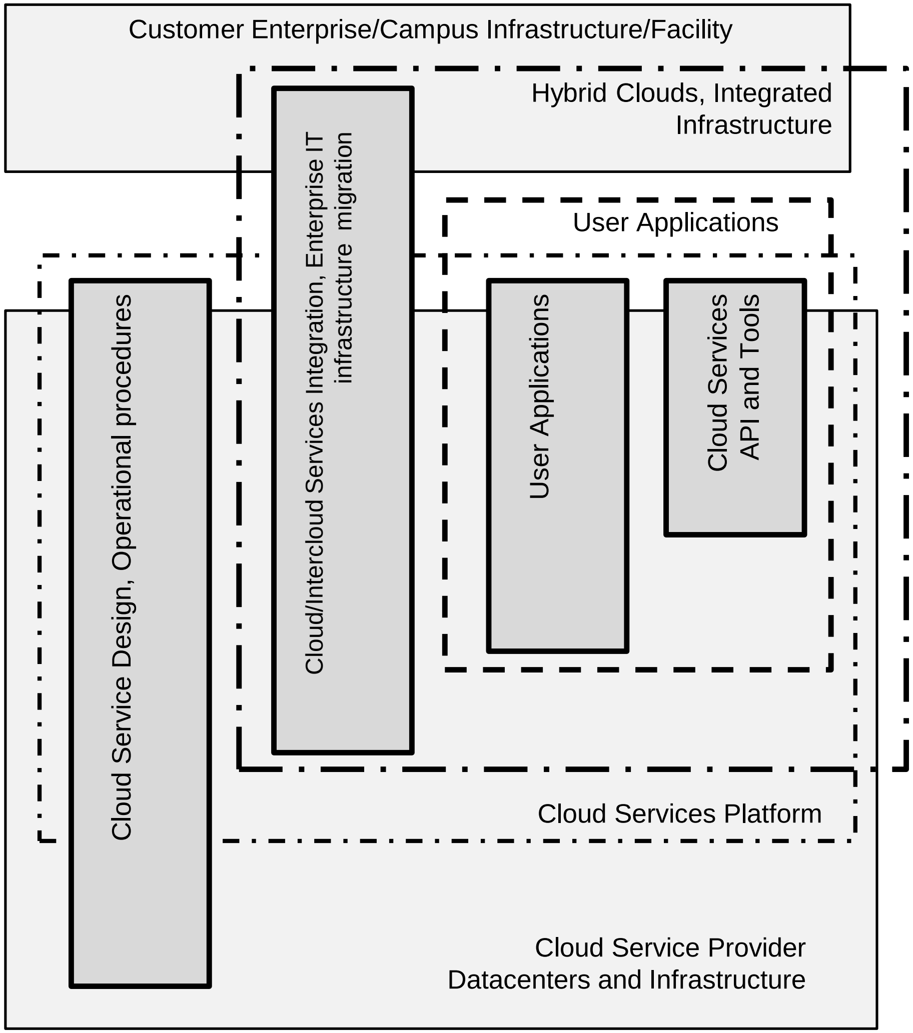
Evaluate (Evaluation)

Create (Synthesis)

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

Taxonomy Cognitive Domain [3]

- Knowledge
- Comprehension
- Application
- Analysis
- Synthesis
- 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

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

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)

Applying Andragogy to Self-Education and Online Training - Problems

- Andragogy concept is widely used in on-line education but
  - 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