Big Data course and Learning Model for Online education (LMO)

at the Laureate Online Education (University of Liverpool)

Yuri Demchenko (University of Amsterdam)
Emanuel Gruengard (Laureate Online Education)

RDA EDISON Workshop

21 September 2014, Amsterdam

Outline

- Need for professional education in Big Data
- Big Data definition and Big Data Architecture Framework (BDAF)
- Common Body of Knowledge in Big Data
- Collaborative Online Learning Model Principles at Laureate Online Education (LOE)
- Big Data and Data Analytics Course
- Bloom's Taxonomy and Andragogy
- Summary and next steps

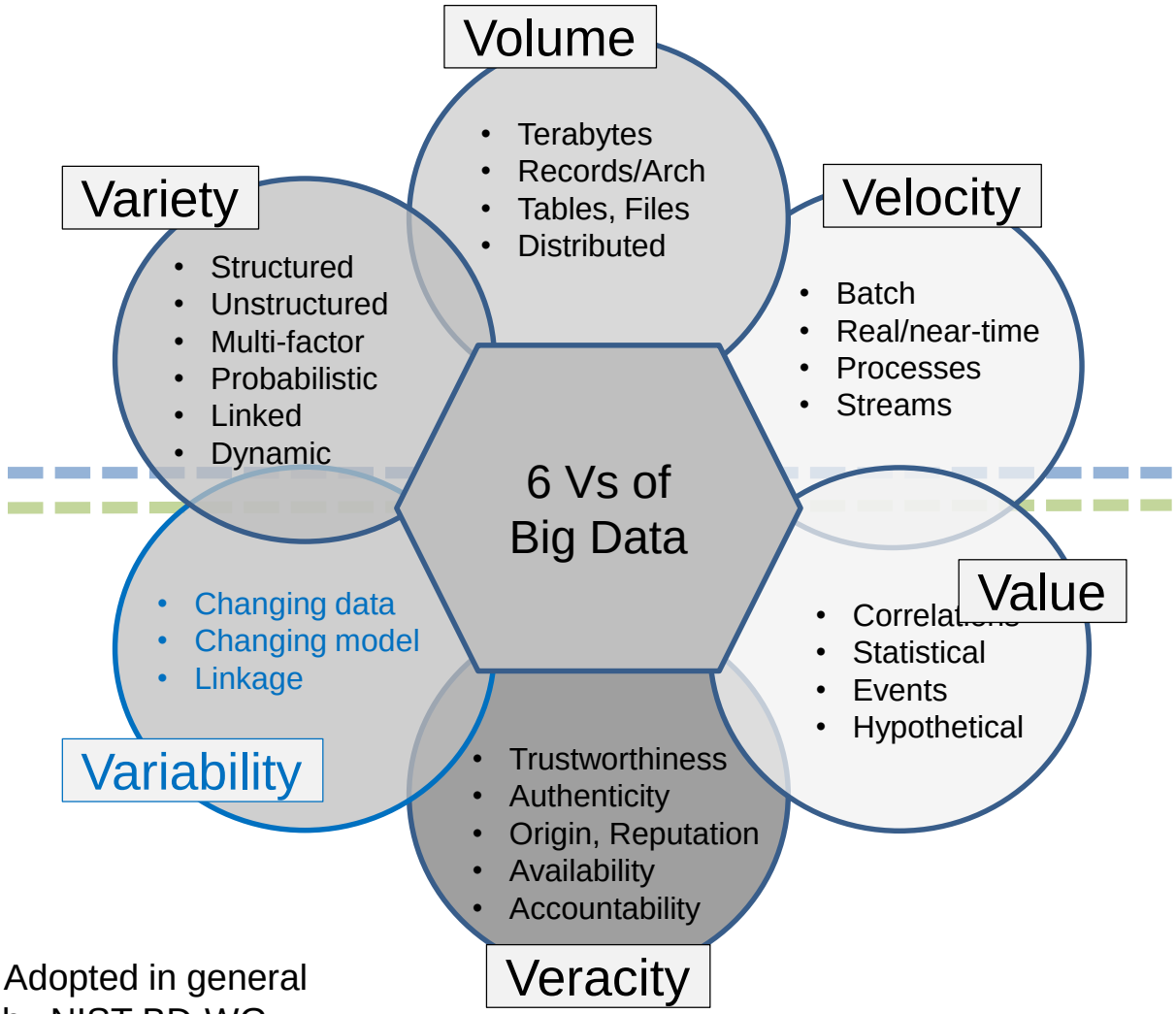
Professional Education Objectives

An effective professional education needs to provide for the professional level of knowledge to achieve the following

1) Master basic concepts and major application areas
2) Compare similar concepts (and concepts inter-relation) and alternatives, as well as application specific areas
3) Appraise basic technologies and their relation to the basic concepts

- Challenges due to Big Data is very wide technology domain
  - Comparing to still narrower Cloud Computing
- New types of skills in Big Data
  - Analytics and research methods

Improved: 6 (5+1) V's of Big Data



Adopted in general by NIST BD-WG

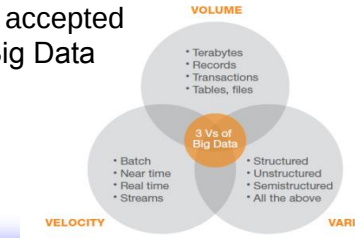
Generic Big Data Properties

- Volume
- Variety
- Velocity

Acquired Properties (after entering system)

- Value
- Veracity
- Variability

Commonly accepted 3V's of Big Data



Big Data Definition: From 6V to 5 Parts (1)

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

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

(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 6V to 5 Parts (2)

Refining Gartner definition

"Big data is (1) high-volume, high-velocity and high-variety information assets that demand (3) cost-effective, innovative forms of information processing for (5) enhanced insight and decision making"

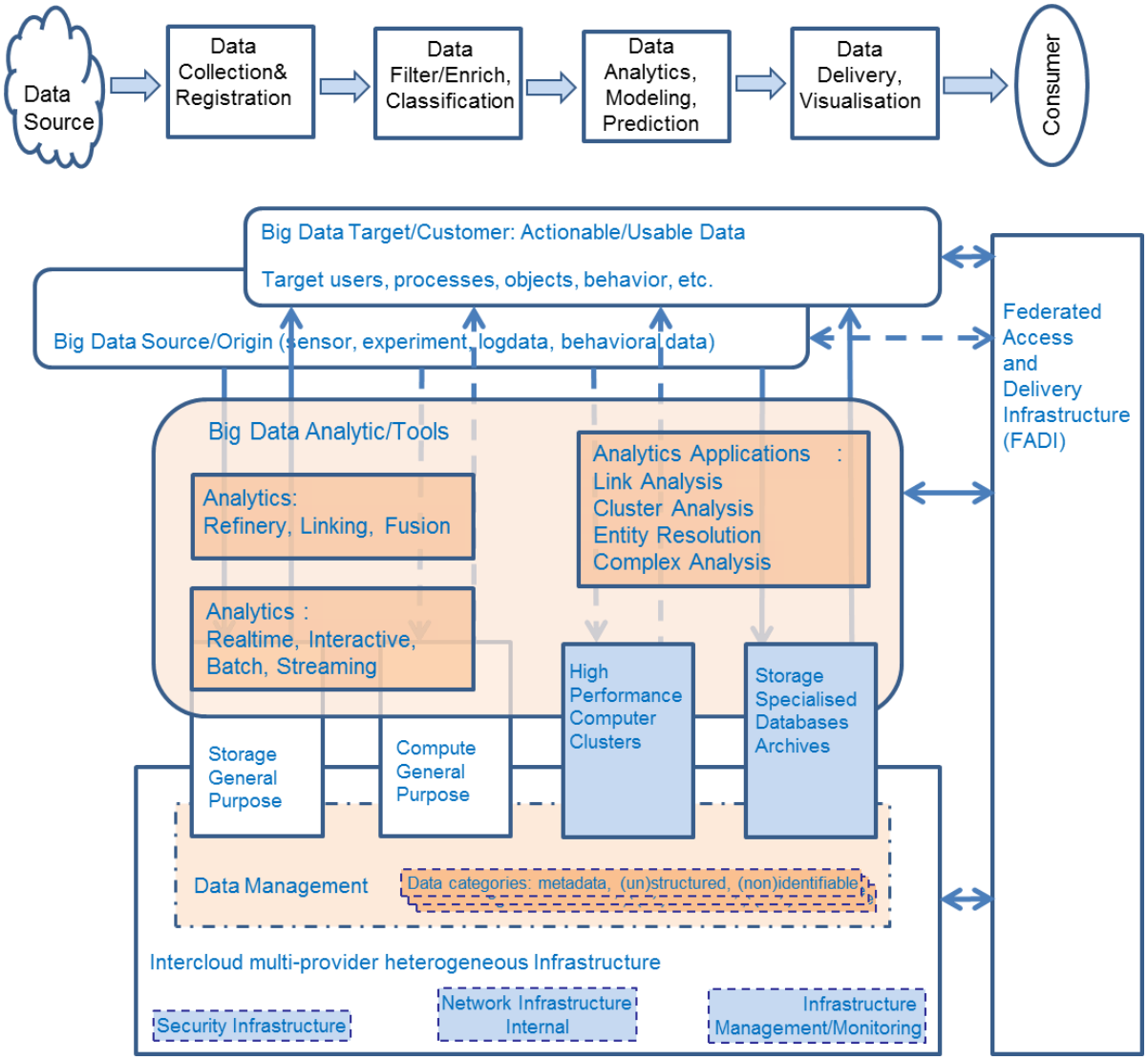
- Big Data (Data Intensive) Technologies are targeting to process (1) high-volume, high-velocity, high-variety data (sets/assets) to extract intended data value and ensure high-veracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (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

Big Data Architecture Framework (BDAF)

- (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 Infrastructure and Analytics Tools



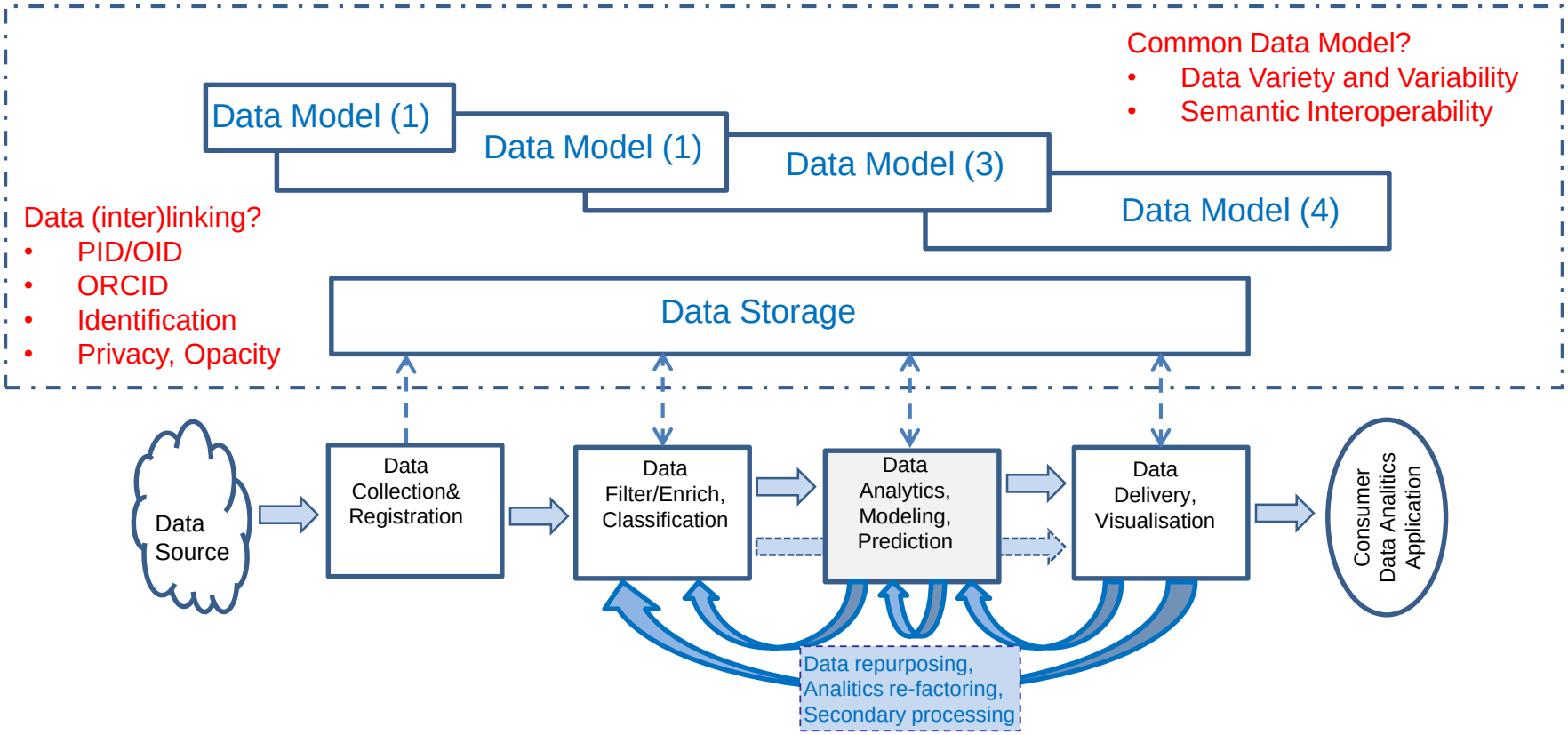
Big Data Infrastructure

- Heterogeneous multi-provider inter-cloud infrastructure
- Data management infrastructure
- Collaborative Environment (user/groups managements)
- Advanced high performance (programmable) network
- Security infrastructure

Big Data Analytics

- High Performance Computer Clusters (HPCC)
- Analytics/processing: Real-time, Interactive, Batch, Streaming
- Big Data Analytics tools and applications

Data Lifecycle/Transformation Model

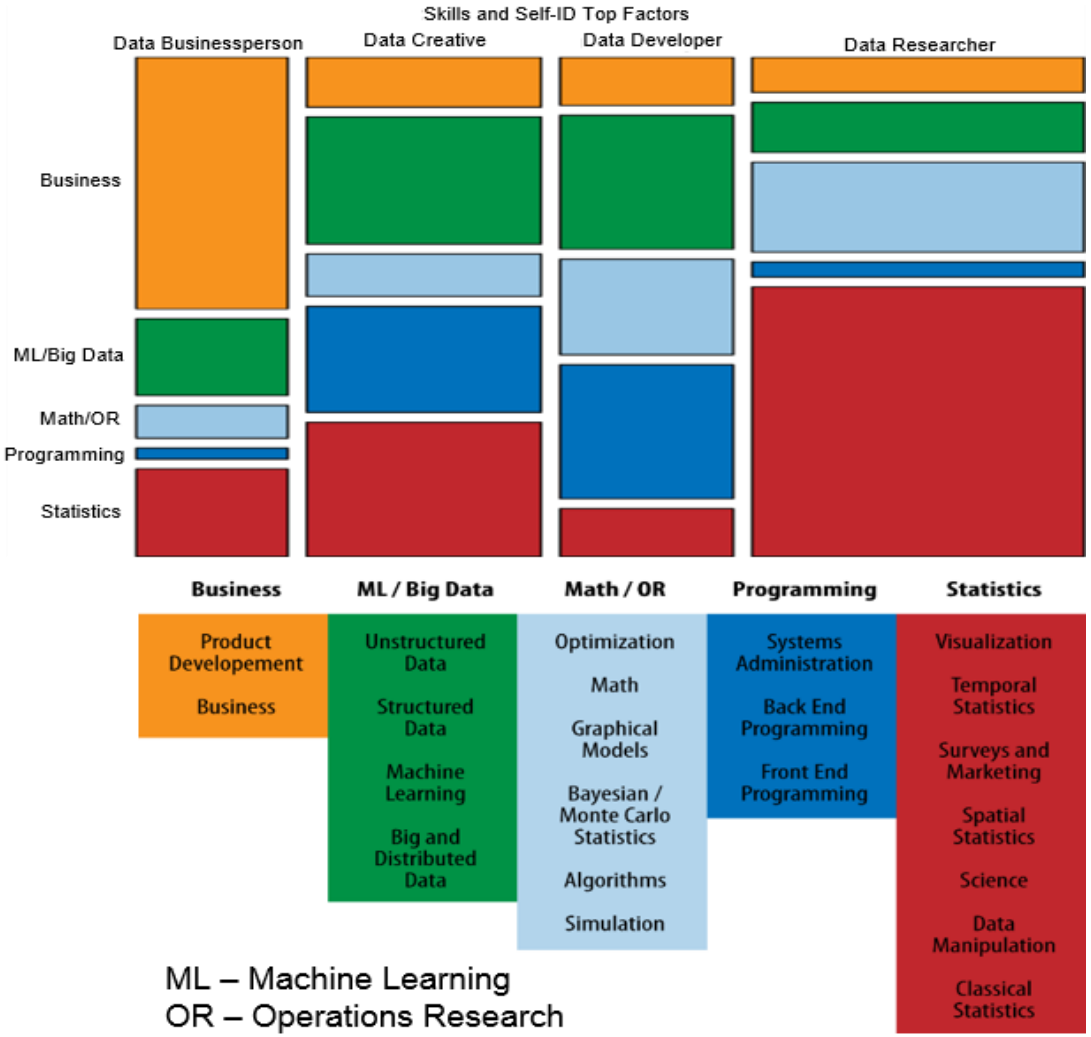


- Does Data Model changes along lifecycle or data evolution?
- Identifying and linking data
  - Persistent identifiers
  - Data ownership
  - Traceability vs Opacity
  - Referral integrity

Big Data and Data Science Skill Taxonomy

- Data Science Competencies Taxonomy by HPC University http://hpcuniversity.org/educators/competencies/
  - Undergraduate Level Computational Science Competencies
  - Graduate Level Computational Science Competencies
  - Basic Data Driven Science Competencies
  - Advanced Data Driven Science Competencies
- Analysing the Analysers. O'Reilly Strata Survey – Harris, Murphy & Vaisman, 2013
- The task of RDA IG on Education and Training

Analysing the Analysers. O'Reilly Strata Survey (2013)



Common Body of Knowledge (CBK) in Big Data and Data Intensive Technologies

CBK refers to several domains or operational categories into which Big Data theory and practices breaks down

- The scope is very wide, need to combine few previously not connected domains
- This is one of attempts verified by practical course development

CBK Big Data and Data Intensive Technologies

1. Big Data Definition and Big Data Architecture Framework, Data driven and data centric applications model, Stakeholders and Roles
2. Big Data use cases and application domains taxonomy and requirements, Big Data in industry and science
3. Data structures, SQL and NoSQL databases
4. Data Analytics Methods and Tools, Knowledge Presentation
5. Big Data Management and curation, Big Data Lifecycle, Data Preservation and Sharing, Enterprise Data Warehouses, Agile Data Driven Enterprise
6. Cloud based Big Data infrastructure and computing platforms, Data Analytics application and new Data Scientist skills required
7. Computing models: High Performance Computing (HPC), Massively Parallel Computing (MPP), Grid, Cluster Computing
8. Big Data Security and Privacy, Certification and Compliance

Big Data Course at LOE/UoL Structure

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: Big Data Infrastructure components and platforms, Enterprise Data Warehouses, MapReduce and Hadoop, distributed file systems and database architectures, data structures, NoSQL databases.

Seminar 4: Big Data analytic techniques, introduction to RapidMiner. Statistical techniques for modeling data.

Seminar 5: Processes behind Big Data Analytics: Rule Extraction Algorithms and Cluster Analysis, Decision tree induction.

Seminar 6: Classification and forecasting techniques: Machine Learning, Neural Networks and Support Vector Machines/ Measurement techniques: Receiver Operating Curves and Gains Charts.

Seminar 7: Big Data Management, Enterprise Data Warehouses (EDW) and emerging *Agile Data Driven Enterprise (ADDE)*, Big Data Service and platform providers.

Seminar 8: Big Data Security and Privacy, data centric security models. Big Data privacy issues and regulations, Privacy Enhancement Techniques.

Laureate Online Education (LOE)

- Laureate Online Education (LOE), the online education partner of the University of Liverpool, provides fully online teaching/education environment based on customized Blackboard platform.
- Laureate's courses are designed to push the boundaries of access to higher education from different countries, cultural backgrounds, and for students with varying educational background.
- The common method here is to push students beyond the boundaries of their customary thinking (i.e., to push them to think "outside of the box") and stimulate their self-motivated learning.

Collaborative Online Learning Model Principles

a) Programs and courses are developed with input from nationally- and internationally-recognized Subject-Matter Experts (SME), leading practitioners, associations/professional groups, and international representatives.
   - Educational materials combine strong conceptual foundation, technology basis and applied mechanisms, standardization, best practices and industry implementation.
   - Programs and courses fully leverage technological and media resources to optimize collaboration and communication.

b) Programs and courses are designed to create an inspiring and transforming student experience and promote collaborative student experiences
   - Programs are future-oriented and forward thinking, both in providing course materials that reflects current status and trends in the technology domain, and in facilitating critical and analytical students' thinking.
   - Students are responsible for their learning and they exercise elements of control over their learning environment. They are inspired through opportunities to engage in reflection and critical thinking, to connect theory to practice, their own experience and educational group experience in the weekly classroom discussions.
   - Work on individual and group projects and hands on assignment.

c) Laureate's programs and courses are designed to expose students to diverse ideas, opinions, perspectives, and experiences - both brought by instructors and based on knowledge and experience exchange in the classroom

d) The course undergoes a quality review process that includes a critical reader review and recommendations, and adoption to the common learning model. The quality reviews are continued all along the course "life span".

Difference from campus education

- Top down vs bottom up approach in course development

Bloom's Taxonomy in Online Education

The courses are developed using best practices for online education and applying Bloom's taxonomy with strong emphasis not only on the Cognitive Domain but also on the Affective Domain to facilitate deep and self-motivated learning. This includes the following:

- Present learning tasks in terms of problem solving, not only as demonstration of accumulated knowledge, and encourage multiple approaches to problem solving.
- Provide opportunities for collaboration with others, including: discussions; sharing of experience, perceptions, and alternate viewpoints; and group activities.
- Allow students to draw on their own experience as part of their learning and to incorporate their own goals into the work of the course.

Andragogy in Adult Education

- Andragogy provides effective approach to online higher education. The following principles of andragogy (adult learning) [19, 20] are incorporated:
- Define a rationale for learning and make a case for the value of doing the work.
- Create environments where self-directed skills are nurtured.
- Have different experiences, background, learning styles, motivation, interests, and goals.
- Have a life-centered orientation to learning; motivate to learn the whole course knowledge domain and show relevance to their professional or career needs.
- Instruction should help the students perform tasks, deal with problems, and thrive in real-life situations.
- Rely on the internal motivation factors and provide such motivators as subject mastering satisfaction, knowledge opening their wider vision and general understanding.

Role of Final Dissertation

- The important role belongs to the final dissertation module where the formation of the future specialist is finalized
  - Duration 9 months
  - Supervised by Dissertation Advisor (DA)]
- The students learn the basics of the research methods and apply them to the dissertation development process that includes
  - Hypothesis and hypothesis verifications
  - Research questions
  - Scholarly contribution
  - Solution development and evaluation

Big Data Module Structure

- The module consists of 8 weekly seminars that includes 2 Discussion Questions (DQ), Hands-in Assignment (HA, or homework) and project assignment.
- Each seminar is provided with the Lecture Notes and textbook reading assignment.
  - There are no synchronous lectures which makes also possible education delivery to countries and to students with low Internet connectivity, as well as bypassing time zone issues.
  - Recorded lectures and accompanying videos are planned for the future.
- Discussion questions and asynchronous discussion are the main form of educational activity.
  - DQ answers are submitted to the discussion forum and the students are required to contribute to the discussion.
  - The students benefit from the knowledge and experience sharing during discussion and learn how to defend own answer.
  - Instructor plays a role of moderator and the students' knowledge and activities assessor.
- Discussion questions are designed in such a way that to stimulate the students' higher cognitive activities starting from the basic literature search to analyzing and evaluating collected and their application to problem solving.
  - Practical use of Bloom's
- Group project and hands on assignments

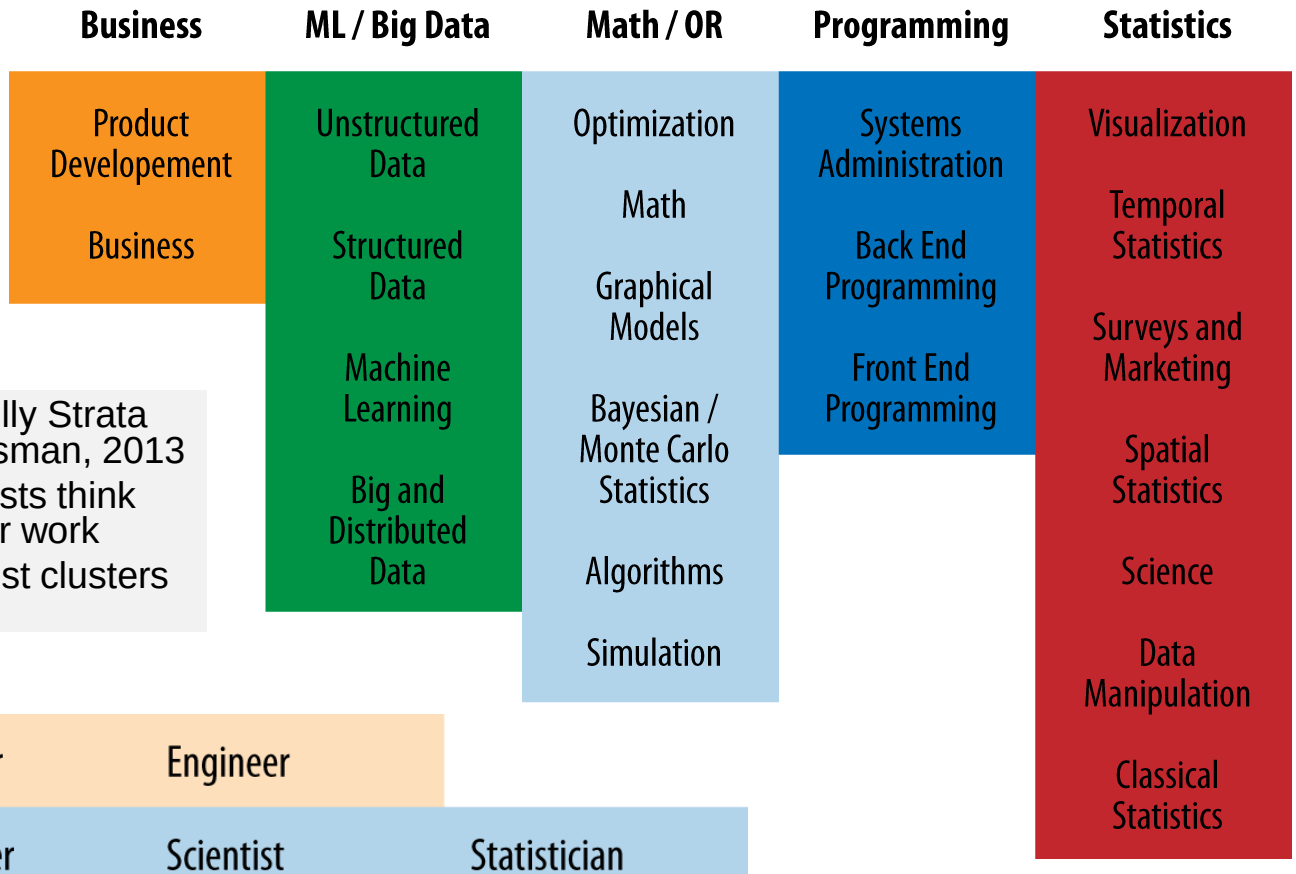
Lessons Learnt

- No books to cover more than 1/3 of the course
  - Plenty of Data Analytics and Machine Learning
  - No Big Data or Scientific Data Infrastructure
    - Some books Data Warehouse, Scientific/HPC computing, NoSQL databases
- Potential use of Cloud Computing and Big Data platforms on clouds is promising but not for such wide students background like at LOE
- Instructor training for online courses guiding and moderation
- Use of MOOC or video lectures – to be considered
- Time to develop good course

Summary and Further Development

- Develop Data Science (Big Data) Infrastructure course for campus education

Strata Survey Skills and Data Scientist Self-ID



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

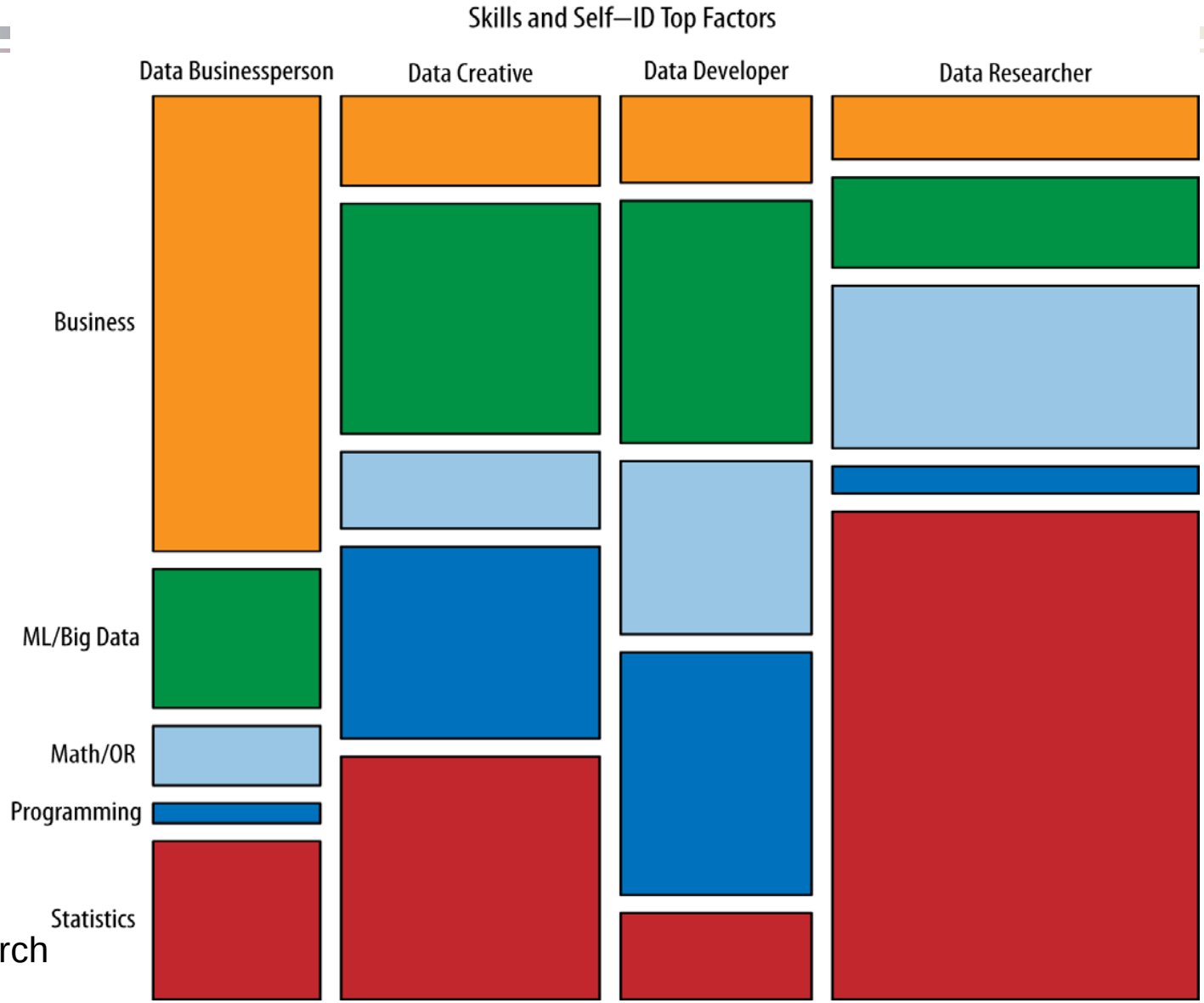
Business

ML/BigData

Math/OR

Programming

Statistics



ML – Machine Learning

OR – Operations Research

Bloom's Taxonomy – Cognitive Activities Example Cloud Computing

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.**

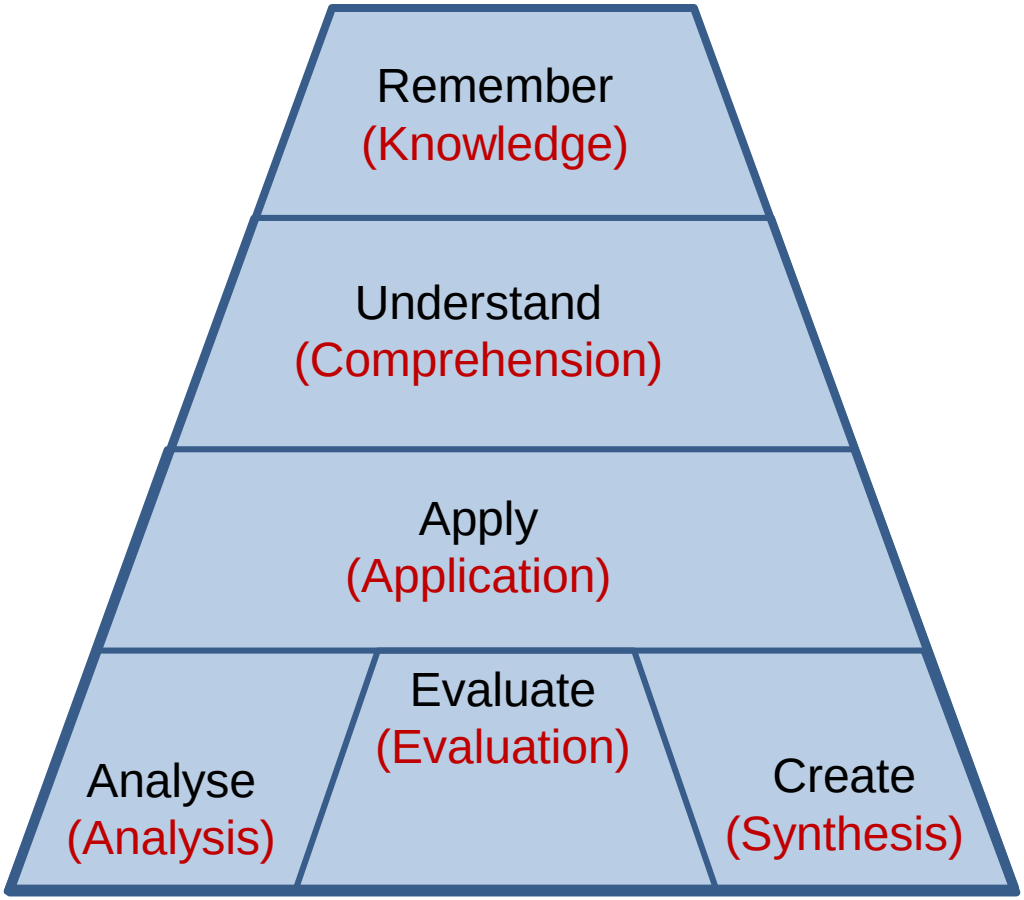
Evaluation

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



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