Education for Data Intensive Science and the EDISON project:

Role of skills management and capacity building for sustainable EOSC development

Yuri Demchenko, EDISON Project
University of Amsterdam

EGI + INDIGO Conference

11 May 2017, Catania, Italy

EDISON – **E**ducation for **D**ata **I**ntensive **S**cience to **O**pen **N**ew science frontiers

Grant 675419 (INFRASUPP-4-2015: CSA)

Outline

- EOSC HLEG Report and Core Data skills gap
  - Bridging cultures between Science and e-Infrastructure
  - Need for conceptual approach to address EOSC challenge of core data experts/skills gap
- EDISON Data Science Framework (EDSF)
  - From Data Science Competences to Body of Knowledge and Model Curriculum
  - New organisational role and Data Science Professional profiles
- Wide spectrum of activities and initiatives worldwide to establish Data (Science) professions family
  - BHEF, DARE/APEC, IEEE/ACM

UvA Universiteit van Amsterdam

University of Southampton

inmark europa

University of Stavanger

EGI

ENGINEERING

FTK Forschungsinstitut für Telekommunikation und Kooperation

Recent European Commission Initiatives 2016

Digitalising European Industry: Reaping the full benefits of a **Digital Single Market**. COM(2016) 180 final, Brussels, 19.4.2016

- The need for new multidisciplinary and digital skills in particular Data Scientist
  - Expected rapidly growing demand will lead to more than 800 000 unfilled vacancies by 2020

European Cloud Initiative - Building a competitive data and knowledge economy in Europe, COM(2016) 178 final, Brussels, 19.4.2016

- **European Open Science Cloud (EOSC)** and European digital research and data infrastructure
  - To offer 1.7 million European researchers and 70 million professionals in science and technology open and seamless services for ***storage, management, analysis and re-use*** of research data
- Address growing demand and shortage of data-related skills

A New Skills Agenda for Europe, COM(2016) 381 final Brussels, 10.6.2016

- Addresses the need for digital and complementary skills, ensure young talents flow into data driven research and industry
- Launched **Digital Skills and Jobs Coalition** (1st December 2016, Brussels) to develop comprehensive national digital skills strategies by mid-2017

HLEG report on European Open Science Cloud (October 2016) – Demand for Core Data Expertise

Realising the European Open Science Cloud. First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud, October 2016

- Recommendation: Allocate 5% grant funding for Data management and preservation
- Estimation: More than 80,000 data stewards to serve 1.7 mln scientists in Europe (1 per every 20 scientists)
- Industry: IDC Report on European Data Market (2015)
  - Number of data workers 6.1 mln (2014) - increase 5.7% from 2013
  - Average number of data workers per company 9.5 - increase 4.4%
  - Gap between demand and supply 509,000 (2014) or 7.5%
- Core data experts need to be trained and their career perspective improved

HLEG EOSC Report Essentials – **Core Data Experts**

- **Core Data Experts** is a new class of colleagues with core scientific professional competencies and the communication skills to fill the gap between the two cultures.
  - **Core data experts** are neither computer savvy research scientists nor are they hard-core data or computer scientists or software engineers.
  - They should be technical data experts, though proficient enough in the content domain where they work routinely from the very beginning (experimental design, proposal writing) until the very end of the data discovery cycle
  - Converge two communities:
    - Scientists need to be educated to the point where they hire, support and respect Core Data Experts
    - Data Scientists (Core Data Experts) need to bring the value to scientific research and organisations
- Implementation of the EOSC needs to include instruments to help train, retain and recognise this expertise,
  - In order to support the 1.7 million scientists and over 70 million people working in innovation.

EOSC Report Recommendations – Implementation on training and skills

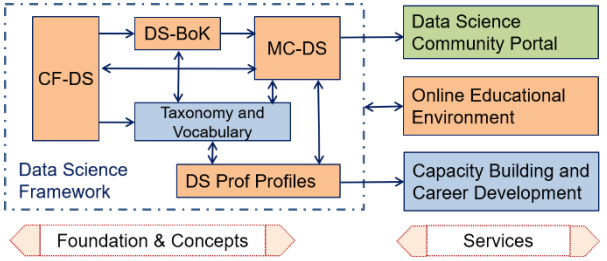
- **I2.1: Set initial guiding principles to kick-start the initiative as quickly as possible. -> Bridge two cultures/communities**
  - A first cohort of core data experts should be trained to translate the needs for data driven science into technical specifications to be discussed with **hard-core data scientists and engineers**.
  - This new class of core data experts will also help translate back to the **hard- core scientists** the technical opportunities and limitations
- **I3: Fund a concerted effort to develop core data expertise in Europe.**
  - Substantial training initiative in Europe to locate, create, maintain and sustain the required core data expertise.
  - **By 2022, to train (hundreds of thousands of) certified core data experts** with a demonstrable effect on ESFRI/e-INFRA activities and prospects for long-term sustainability of this critical human resource
    - Consolidate and further develop assisting material and tools for Data Management Plans and Data Stewardship plans (including long-term preservation in FAIR status)
- **I7: Provide a clear operational timeline to deal with the early preparatory phase of the EOSC.**
  - **Define training needs for the necessary data expertise and draw models for the necessary training infrastructure**

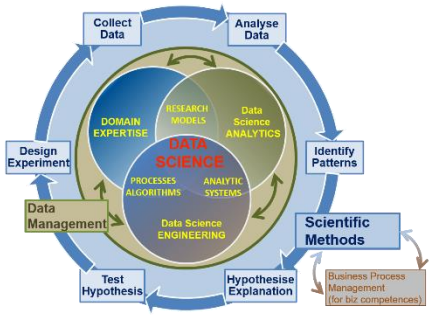
Approach

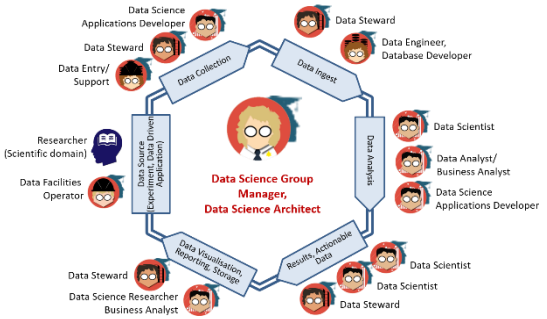
- Task is not for one community or one project
  - Need collaboration between different stakeholders and communities: academia, research, industry, public sector
- Task is not for science or RI only in isolation from industry and academia
- Needs strong conceptual approach
  - Use science to solve the problems of science
- Standardisation is an important factor of sustainability and development

EDISON Services for Data Science and Skills Management and Tailored Education

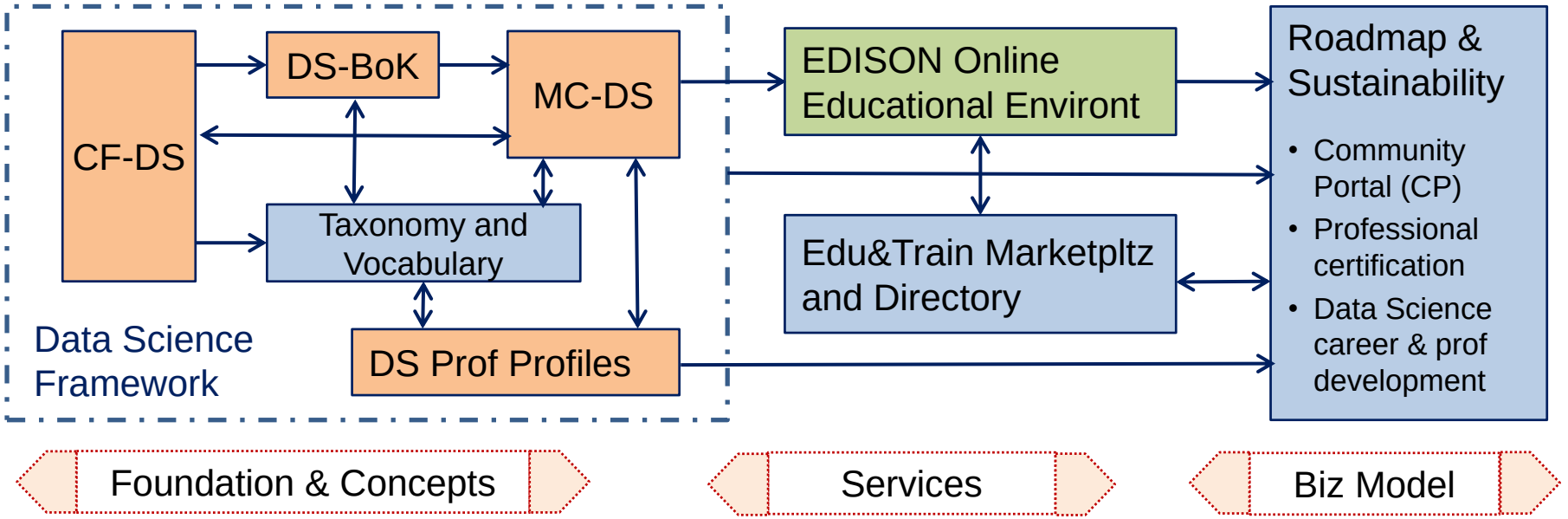
- EDISON Data Science Framework (EDSF)
  - Compliant with EU standards on competences and professional occupations e-CFv3.0, ESCO
  - Customisable courses design for targeted education and training
- Skills development and career management for Core Data Experts and related data handling professions
- Capacity building and Data Science team design
- Academic programmes and professional training courses (self) assessment and design
- EU network of Champion universities pioneering Data Science academic programmes
- Engagement in relevant RDA activities and groups
- Cooperation with International professional organisations IEEE, ACM, BHEF, APEC (AP Economic Cooperation)





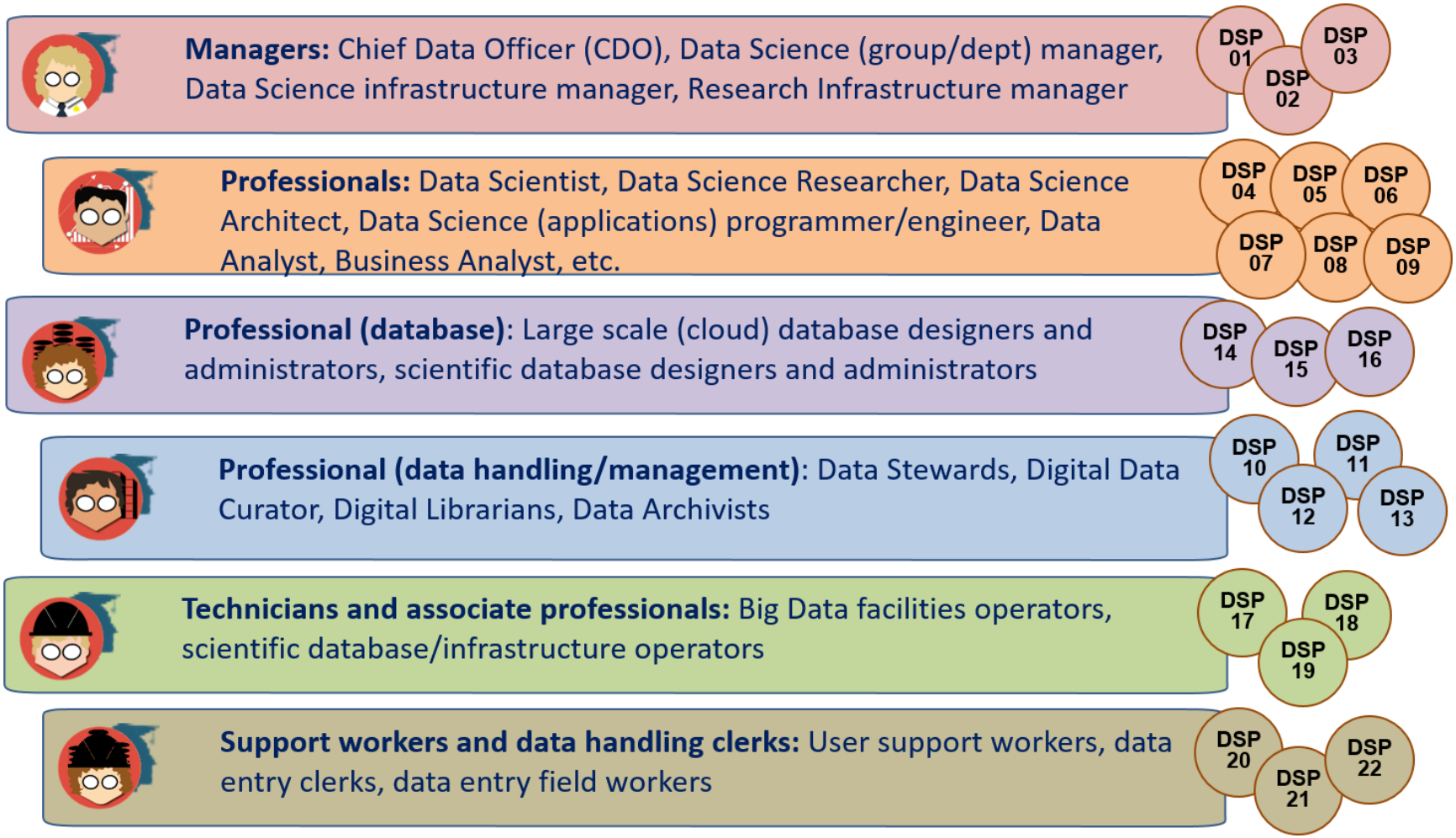


EDISON Data Science Framework (EDSF) Release 1 (October 2016)



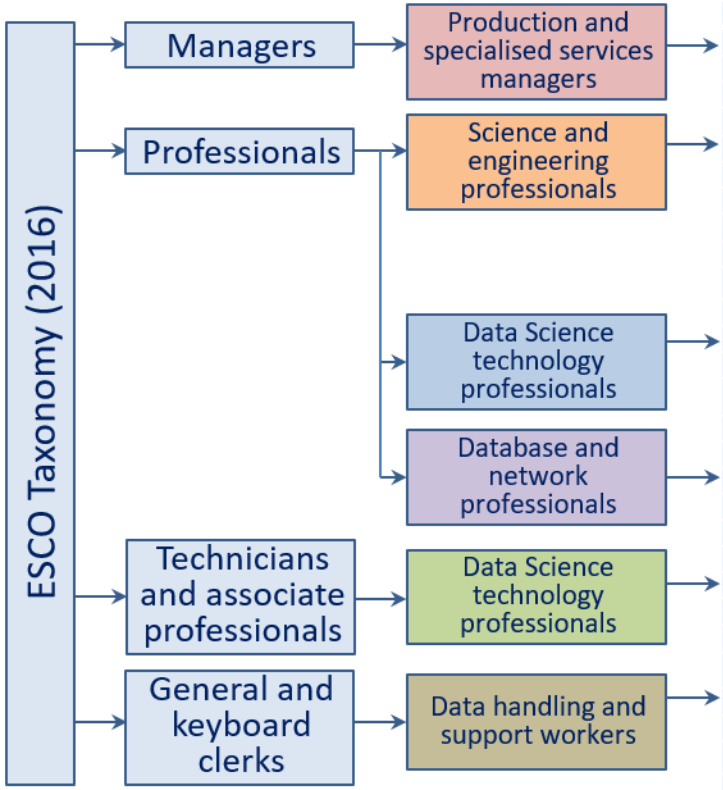
- EDISON Framework components
  - CF-DS – Data Science Competence Framework
  - DS-BoK – Data Science Body of Knowledge
  - MC-DS – Data Science Model Curriculum
  - DSP – Data Science Professional profiles
  - Data Science Taxonomies and Scientific Disciplines Classification
  - EOEE - EDISON Online Education Environment

Data Science Professions Family



Icons used: Credit to [ref] https://www.datacamp.com/community/tutorials/data-science-industry-infographic

DSP Profiles mapping to ESCO Taxonomy High Level Groups



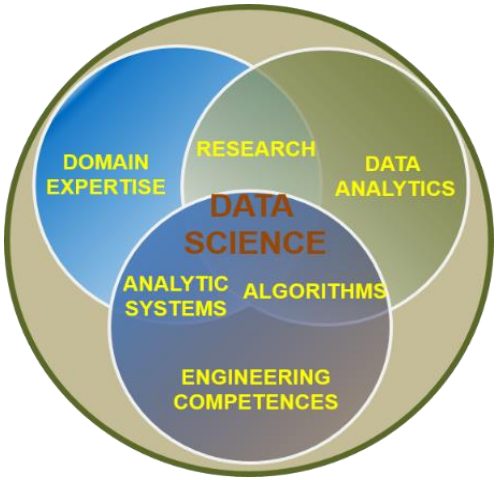
Profile ID	Data Science Profile title	DSDA	DSDM	DSENG	DSRM	DSDK
Data Science Services/Infrastructure Managers						
DSP01	Data Science (group) Manager	3	4	3	3	2
DSP02	Data Science Infrastr Manager	2	4	4	2	2
DSP03	Research Infrastructure Manager	2	4	4	3	2
Data Scince Professionals						
DSP04	Data Scientist	5	3	4	5	3
DSP05	Data Science Researcher	4	3	2	5	4
DSP06	Data Science Architect	4	3	5	3	3
DSP07	Data Science Applic Programmer	4	2	5	3	4
DSP08	Data Analyst	5	3	3	3	4
DSP09	Business Analyst	5	3	3	4	5
Data handling professionals not elsewhere classified						
DSP10	Data Stewards	3	5	3	3	3
DSP11	Digital data curator	1	5	2	2	3
DSP12	Digital Librarians	2	5	2	2	3
DSP13	Data Archivists	1	5	1	1	3
Database and network professionals not elsewhere classified						
DSP14	Large scale database designer	2	4	4	3	3
DSP15	Large scale database admin	2	4	3	2	3
DSP16	Scientific database administrator	2	4	3	2	3
Data Infrastructure engineers and technicians						
DSP17	Big Data facilities Operator	1	4	4	2	3
DSP18	Large scale data storage operator	1	4	3	1	1
DSP19	Scientific database operator	1	4	3	2	3
Data and information entry and access						
DSP20	Data entry/access worker		2	1		2
DSP21	Data entry field workers		2	1		2
DSP22	User support data services		3	2		2

- DSP Profiles mapping to corresponding CF-DS Competence Groups
  - Relevance level from 5 – maximum to 1 – minimum

Data Scientist definition

Based on the definitions by NIST Big Data WG (NIST SP1500 - 2015)

- *A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in **business needs, domain knowledge, analytical skills, and programming and systems engineering expertise** to manage the end-to-end scientific method process through each stage in the **big data lifecycle***
  - *… Till the delivery of an **expected scientific and business value** to science or industry*
- *Other definitions to admit such features as*
  - Ability to solve variety of business problems
  - Optimize performance and suggest new services for the organisation
  - Develop a special mindset and be statistically minded, ***understand raw data** and **"appreciate data as a first class product"***



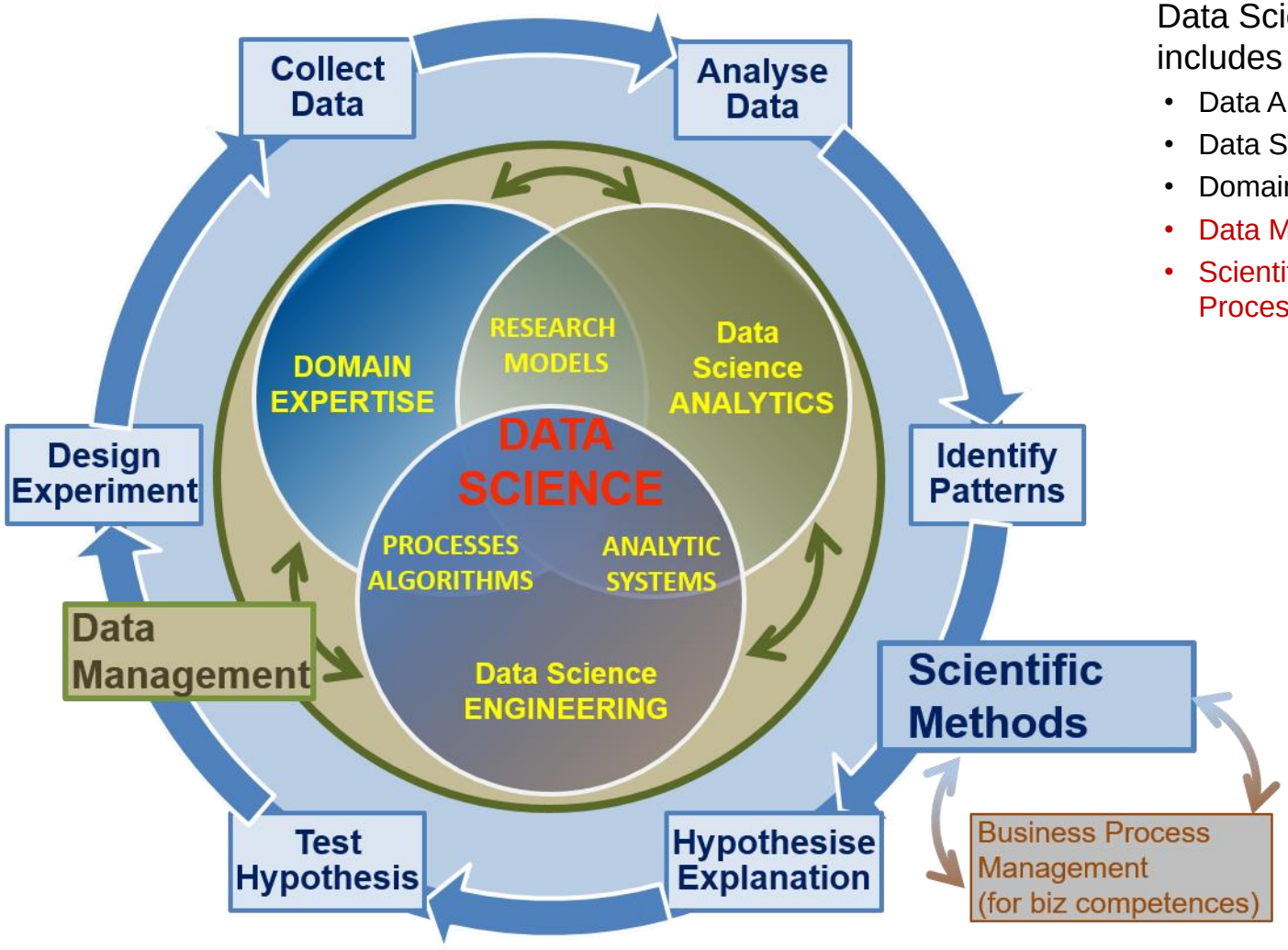
[ref] Legacy: NIST BDWG definition of Data Science

- ***Data science** is the empirical synthesis of actionable knowledge and technologies required to handle data from raw data through the complete data lifecycle process.*
- ***Big Data** is the technology to build system and infrastructures to process large volume of structurally complex data in a time effective way*

Identified Data Science Competence Groups

- Core Data Science competences/skills groups
  - **Data Science Analytics** (including Statistical Analysis, Machine Learning, Business Analytics)
  - **Data Science Engineering** (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
  - **Domain Knowledge and Expertise** (Subject/Scientific domain related)
- EDISON identified 5 core competence groups demanded by organisations
  - **Data Management, Curation, Preservation**
  - **Scientific or Research Methods and/vs Business Processes/Operations**
- Other skills commonly recognized aka “soft skills” or “21st Century Skills”
  - Inter-personal skills and team work, cooperativeness
- Important aspect of integrating Data Scientist (team) into organisation structure
  - General Data Science (and Data) **literacy** for all involved roles and management
  - ***Role of Data Scientist: Provide a kind of literacy advice and guidance to organisation***

Data Science Competence Groups - Research



Data Science Competence includes 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- Data Management
- Scientific Methods (or Business Process Management)

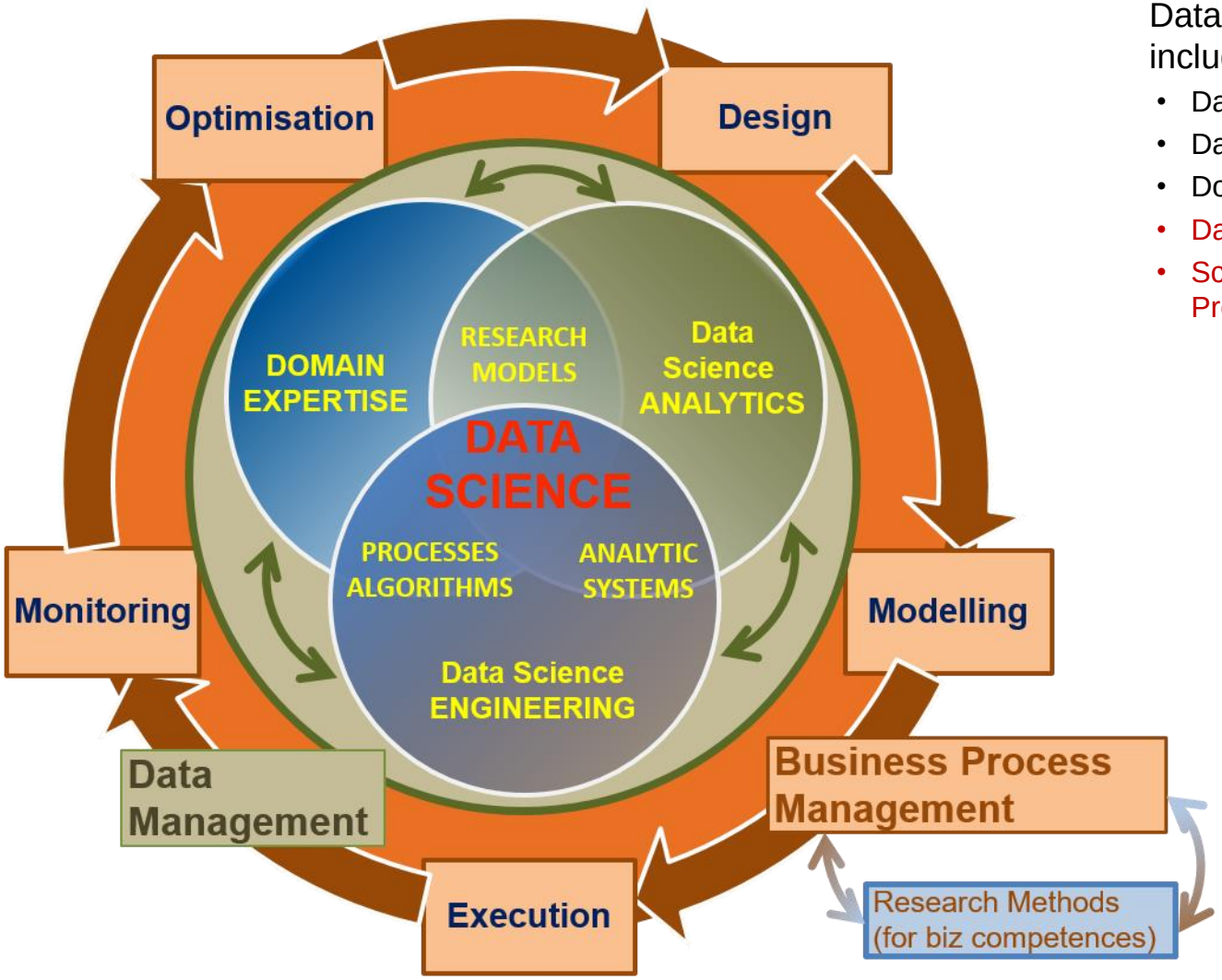
Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

Business Operations

- Operations Strategy
- Plan
- Design & Deploy
- Monitor & Control
- Improve & Re-design

Data Science Competences Groups – Business



Data Science Competence includes 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- Data Management
- Scientific Methods (or Business Process Management)

Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

Business Process Operations/Stages

- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design

Identified Data Science Competence Groups

	Data Science Analytics (DSDA)	Data Management (DSDM)	Data Science Engineering (DSENG)	Research/Scientific Methods (DSRM)	Data Science Domain Knowledge, e.g. Business Processes (DSDK/DSBPM)
0	Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management	Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals	Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
1	**DSDA01** Use predictive analytics to analyse big data and discover new relations	**DSDM01** Develop and implement data strategy, in particular, Data Management Plan (DMP)	**DSENG01** Use engineering principles to design, prototype data analytics applications, or develop instruments, systems	**DSRM01** Create new understandings and capabilities by using scientific/ research methods or similar domain related development methods	**DSBPM01** Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
2	**DSDA02** Use statistical techniq to deliver insights	**DSDM02** Develop data models including metadata	**DSENG02** Develop and apply computational solutions	**DSRM02** Direct systematic study toward a fuller knowledge or understanding of the observable facts	**DSBPM02** Participate strategically and tactically in financial decisions
3	**DSDA03** Develop specialized …	**DSDM03** Collect integrate data	**DSENG03** Develops specialized tools	**DSRM03** Undertakes creative work	**DSBPM03** Provides support services to other
4	**DSDA04** Analyze complex data	**DSDM04** Maintain repository	**DSENG04** Design, build, operate	**DSRM04** Translate strategies into actions	**DSBPM04** Analyse data for marketing
5	**DSDA05** Use different analytics	**DSDM05** Visualise cmplx data	**DSENG05** Secure and reliable data	**DSRM05** Contribute to organizational goals	**DSBPM05** Analyse optimise customer relatio

Identified Data Science *Skills/Experience* Groups

- **Group 1: Skills/experience related to competences**
  - Data Analytics and Machine Learning
  - Data Management/Curation (including both general data management and scientific data management)
  - Data Science Engineering (hardware and software) skills
  - Scientific/Research Methods or Business Process Management
  - Application/subject domain related (research or business)
  - Mathematics and Statistics
- **Group 2: Big Data (Data Science) tools and platforms**
  - Big Data Analytics platforms
  - Mathematics & Statistics applications & tools
  - Databases (SQL and NoSQL)
  - Data Management and Curation platform
  - Data and applications visualisation
  - *Cloud based platforms and tools*
- **Group 3: Programming and programming languages and IDE**
  - General and specialized development platforms for data analysis and statistics
- **Group 4: Soft skills or Social Intelligence**
  - Personal, inter-personal communication, team work, professional network

Practical Application of the CF-DS

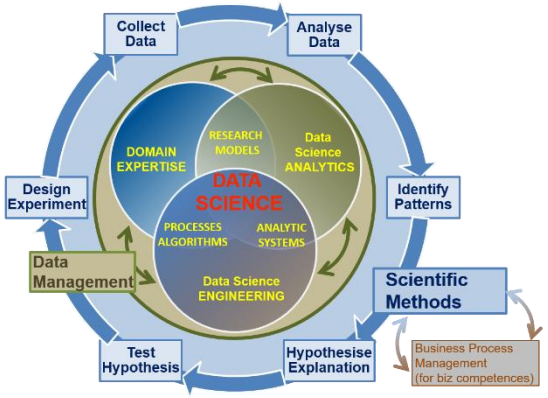
- Basis for the definition of the Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS)
  - CF-DS => Learning Outcomes (MC-DS) => Knowledge Areas (DS-BoK)
  - CF-DS => Data Science taxonomy of scientific subjects and vocabulary
- Data Science professional profiles definition
  - Extend existing EU standards and occupations taxonomies: e-CFv3.0, ESCO, others
- Professional competence benchmarking
  - For customizable training and career development
  - Including CV or organisational profiles matching
- Professional certification
  - In combination with DS-BoK professional competences benchmarking
- Vacancy construction tool for job advertisement (for HR)
  - Using controlled vocabulary and Data Science Taxonomy

Education and Training

- Foundation and methodological base
  - Data Science Body of Knowledge (DS-BoK)
    - Taxonomy and classification of Data Science related scientific subjects
  - Data Science Model Curriculum (MC-DS)
    - Set Learning Units mapped to CF-DS Learning and DS-BoK Knowledge Areas/Units
  - Instructional methodologies and teaching models
- Platforms and environment
  - Virtual labs, datasets, developments platforms
  - Online education environment and courses management
- Services
  - Individual benchmarking and profiling tools (competence assessment)
  - Knowledge evaluation tools
  - Certifications and training for self-made Data Scientists practitioners
  - Education and training marketplace: Courses catalog and repository

Data Science Body of Knowledge (DS-BoK)

DS-BoK Knowledge Area Groups (KAG)



- KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRM: *Scientific/Research Methods group*
- KAG5-DSBP: Business process management group

- Data Science domain knowledge to be defined by related expert groups

Data Science Model Curriculum (MC-DS)

Data Science Model Curriculum includes

- Learning Outcomes (LO) definition based on CF-DS
  - LOs are defined for CF-DS competence groups and for all enumerated competences
- LOs mapping to Learning Units (LU)
  - LUs are based on CCS(2012) and universities best practices
  - Data Science university programmes and courses inventory (interactive) http://edison-project.eu/university-programs-list
- LU/course relevance: Mandatory Tier 1, Tier 2, Elective, Prerequisite
- Learning methods and learning models (in progress)

Example MC-DS Mapping Learning Units to DS-BoK and CCS (2012)

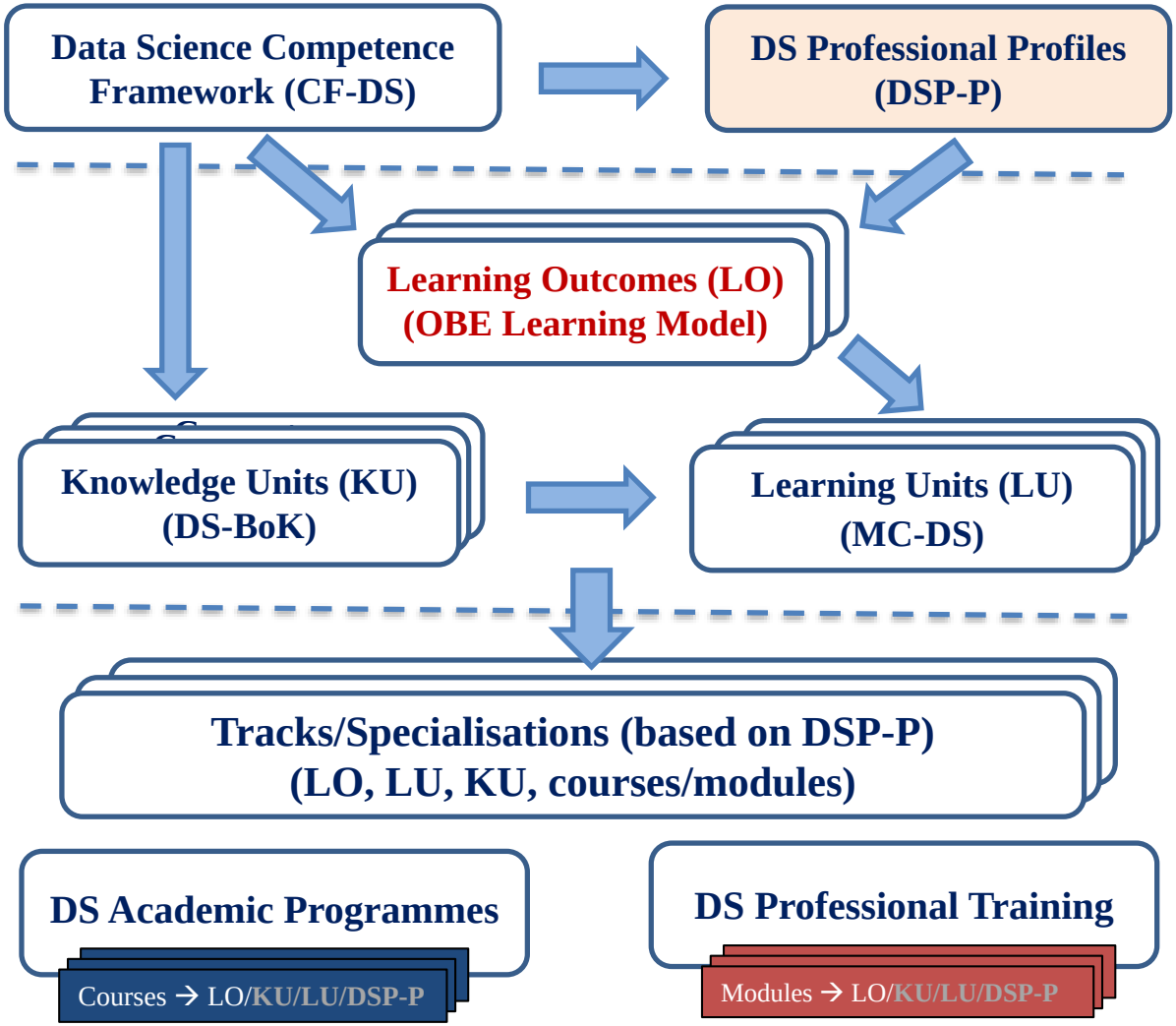
KAG/ LU# *)	Learning Unit (course name) [2]	Type/relevance [3]				Map to DS-BoK, CCS2012 and known BoKs	
		Tier 1	Tier 2	Elective	Pre requisite	CCS2012 based academic subjects	DS-BoK and other BoKs
	Software requirements and design					Extensions are suggested from SWEBOK	SWEBOK selected KAs • Software requirements

KAG/ LU# *)	Learning Unit (course name) [2]	Type/relevance [3]				Map to DS-BoK, CCS2012 and known BoKs	
		Tier 1	Tier 2	Elective	Pre requisite	CCS2012 based academic subjects	DS-BoK and other BoKs
	Information theory					Mathematical analysis	
	Mathematical analysis						
	Extensibility point for adding new courses						
	Artificial Intelligence					Computing methodologies Artificial intelligence	No specific BoK are defined
	Natural Language Processing						
	Knowledge Represer Reasoning						
	Data mining and knc discovery						
	Text analysis, Data n						
	Text analytics includ linguistic, and struct techniques to analys and unstructured da						
	Machine Learning th algorithms						
	Classification metho						

KAG/ LU# *)	Learning Unit (course name) [2]	Type/relevance [3]				Map to DS-BoK, CCS2012 and known BoKs	
		Tier 1	Tier 2	Elective	Pre requisite	CCS2012 based academic subjects	DS-BoK and other BoKs
	Data type registries, PID, metadata					Extended with the general Data Management Knowledge Areas and related academic subjects.	General Data Management KA's Data Lifecycle Management Data archives/storage compliance and certification New KAs to support RDA recommendations and community data management models (Open Access, Open Data, etc) Data type registries, PIDs Data infrastructure and Data Factories TBD – To follow RDA and ERA community developments
	Research data infrastructure, Open Science, Open Data, Open Access, ORCID						
	Extensibility point for adding new courses						
	Research methodology, research cycle					Extended with the general Scientific/Research Methods subjects and related academic subjects.	Suggested KAs to develop DSRM related competences: Research methodology, research cycle (e.g. 4 step model Hypothesis – Research Methods – Artefact –
	Modelling and experiment planning						

- Mapping suggested to ACM CCS2012, DS-BoK and other related BoKs

Outcome Based Educations and Training Model



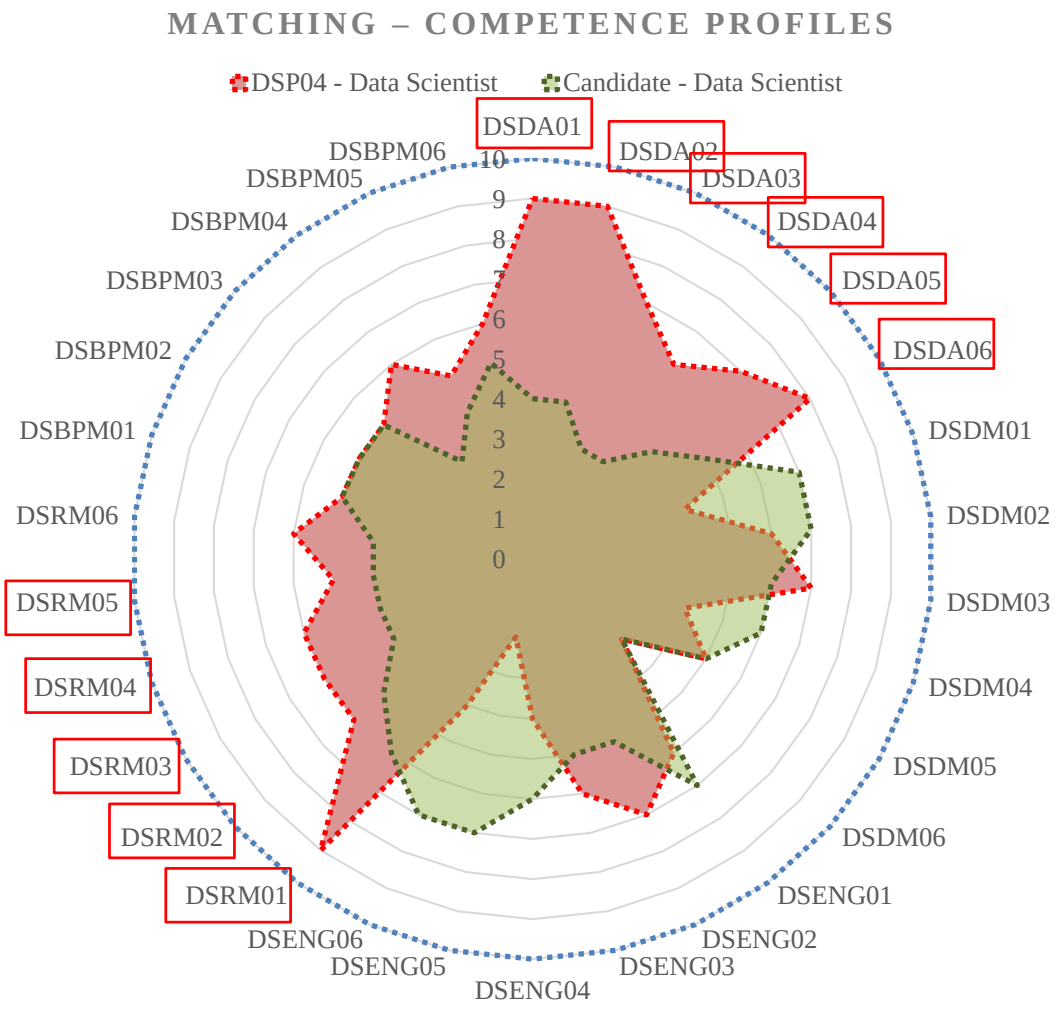
From Competences and DSP Profiles

to Learning Outcomes (LO) and

to Knowledge Unites (KU) and Learning Units (LU)

- EDSF allow for customized educational courses and training modules design

Individual Competences Benchmarking

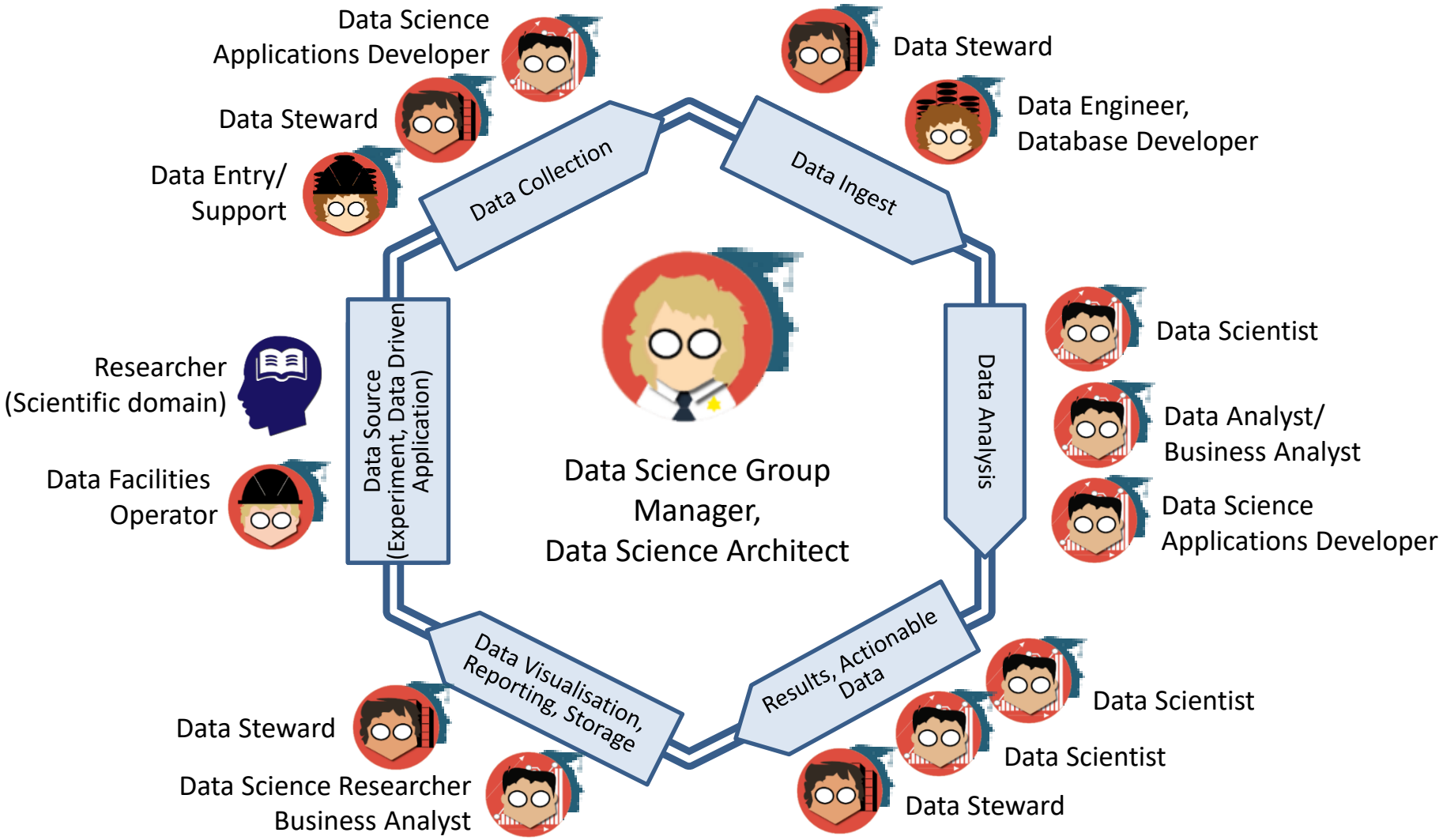


Individual Education/Training Path based on Competence benchmarking

- Red polygon indicates the chosen professional profile: Data Scientist (general)
- Green polygon indicates the candidate or practitioner competences/skills profile
- Insufficient competences (gaps) are highlighted in *red*
  - *DSDA01 – DSDA06 Data Science Analytics*
  - *DSRM01 – DSRM05 Data Science Research Methods*
- Can be use for team skills match marking and organisational skills management

[ref] For DSP Profiles definition and for enumerated competences refer to EDSF documents CF-DS and DSP Profiles.

Building a Data Science Team



Data Science or Data Management Group/Department: Organisational structure and staffing - EXAMPLE

Data Science or Data Management Group/Department

>> Reporting to CDO/CTO/CEO

- Providing cross-organizational services

- (Managing) Data Science Architect (1)
- Data Scientist (1), Data Analyst (1)
- Data Science Application programmer (2)
- Data Infrastructure/facilities administrator/operator: storage, cloud, computing (1)
- Data stewards, curators, archivists (3-5)

Estimated: Group of 10-12 data specialists for research institution of 200-300 research staff.

KAG3-DSDM: *Data Management group: data curation, preservation and data infrastructure*

DM-BoK version 2 “Guide for performing data management” – 11 Knowledge Areas

(1) Data Governance

(2) Data Architecture

(3) Data Modelling and Design

(4) Data Storage and Operations

(5) Data Security

(6) Data Integration and Interoperability

(7) Documents and Content

(8) Reference and Master Data

(9) Data Warehousing and Business Intelligence

(10) Metadata

(11) Data Quality

Other Knowledge Areas motivated by RDA, European Open Data initiatives, European Open Data Cloud

(12) PID, metadata, data registries

(13) Data Management Plan

(14) Open Science, Open Data, Open Access, ORCID

(15) Responsible data use

- Highlighted in red: Considered (Research) Data Management literacy (minimum required knowledge)

Useful links

- EDISON project website http://edison-project.eu/

- EDISON Data Science Framework Release 1 (EDSF)
  http://edison-project.eu/edison-data-science-framework-edsf
  - Data Science Competence Framework
    http://edison-project.eu/data-science-competence-framework-cf-ds
  - Data Science Body of Knowledge
    http://edison-project.eu/data-science-body-knowledge-ds-bok
  - Data Science Model Curriculum
    http://edison-project.eu/data-science-model-curriculum-mc-ds
  - Data Science Professional Profiles
    http://edison-project.eu/data-science-professional-profiles-definition-dsp

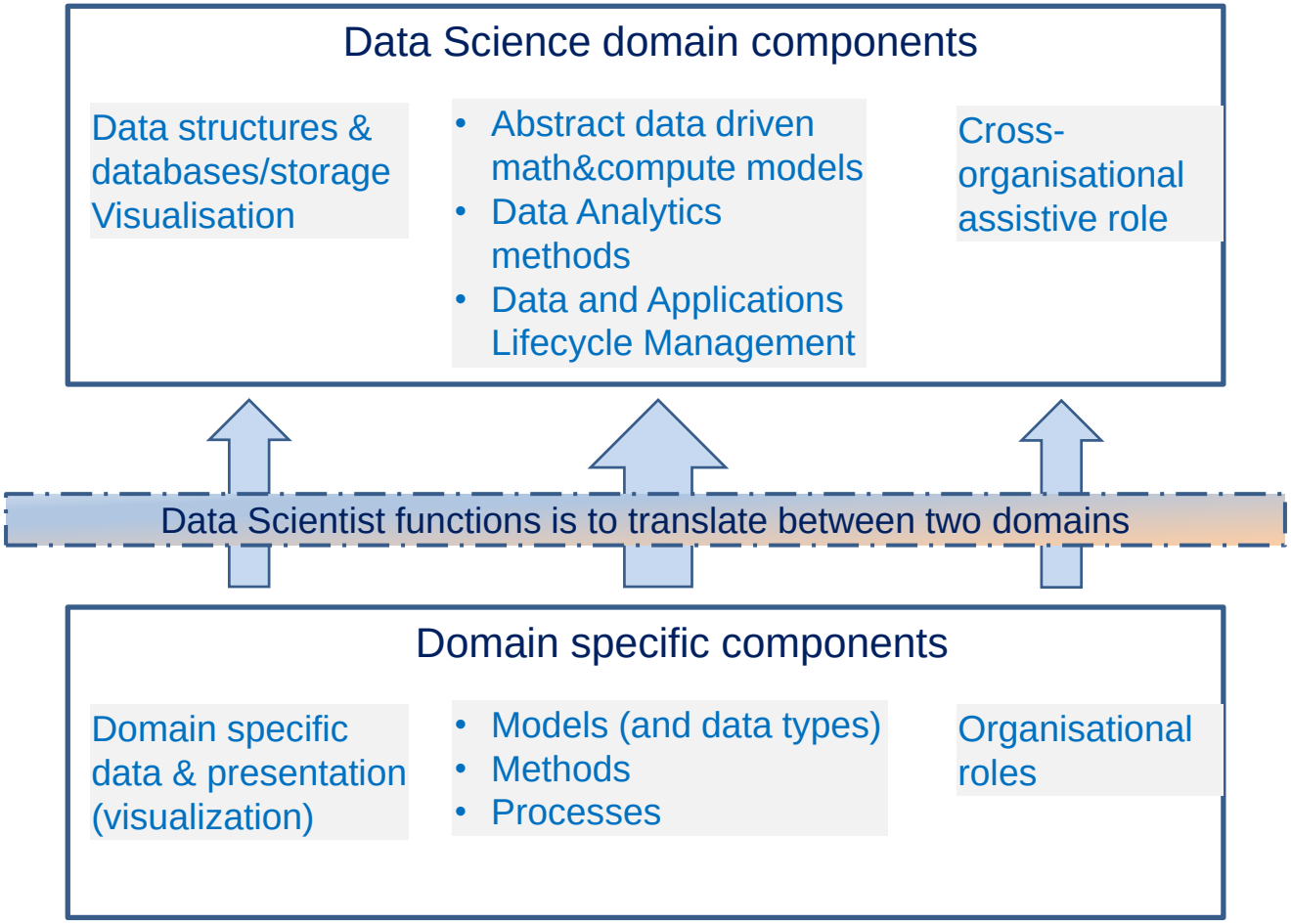
- Survey Data Science Competences: Invitation to participate
  https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession

Data Scientist and Subject Domain Specialist

- **Subject domain components**
  - Model (and data types)
  - Methods
  - Processes
  - Domain specific data and presentation/visualization methods
  - Organisational roles and relations

- **Data Scientist is an assistant to Subject Domain Specialists**
  - Translate subject domain Model, Methods, Processes into abstract data driven form
  - Implement computational models in software, build required infrastructure and tools
  - Do (computational) analytic work and present it in a form understandable to subject domain
  - Discover new relations originated from data analysis and advice subject domain specialist
  - Present/visualise information in domain related actionable way
  - Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data

Data Science and Subject Domains



Data Scientist role is to maintain the Data Value Chain (domain specific):

- Data Integration => Organisation/Process/Business Optimisation => **Innovation**