EDISON Data Science Framework (EDSF) as a basis for Data Science Curriculum Harmonisation, Skills Management and Capacity Building

Yuri Demchenko, EDISON Project
University of Amsterdam

Data Science Curriculum Harmonisation

26 May 2017, Kiev, Ukraine

EDISON
building the data science profession

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

Grant 675419 (INFRASUPP-4-2015: CSA)

Outline

- European Digital Single Market (DSM) and demand for data enabled skills
  - Recent European Commission Initiatives 2016
- EDISON Data Science Framework (EDSF)
  - From Data Science Competences and Skills to Body of Knowledge and Model Curriculum
  - Data Science Profession Profiles family and organisational skills management
- Examples Data Science curricula and EDSF tools
  - Professional issues in Data Science
- Activities and initiatives worldwide to establish Data (Science) professions family
  - BHEF, DARE/APEC, IEEE/ACM
- Rise of a Data Steward: EOSC HLEG Report and Core Data skills gap
  - Need for conceptual approach to address EOSC challenge of core data experts/skills gap
- Summary and discussion

UvA Universiteit van Amsterdam

University of Southampton

inmark europa

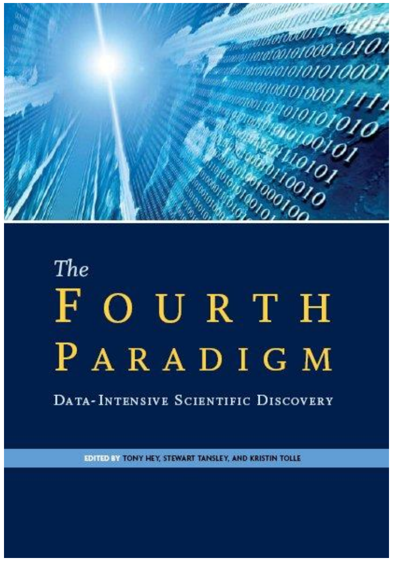
University of Stavanger

egi

ENGINEERING

FTK Forschungsinstitut für Telekommunikation und Kooperation

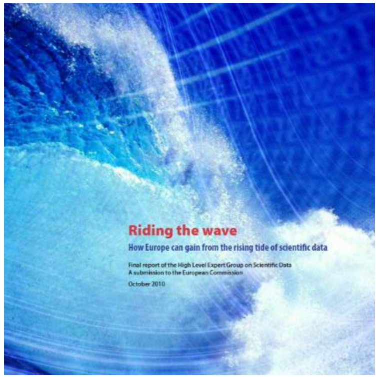
Visionaries and Drivers:
Seminal works, High level reports, Activities


The FOURTH PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

The Fourth Paradigm: Data-Intensive Scientific Discovery.

By Jim Gray, Microsoft, 2009. Edited by Tony Hey, Kristin Tolle, et al.

http://research.microsoft.com/en-us/collaboration/fourthparadigm/


Riding the wave
How Europe can gain from the rising tide of scientific data
Final report of the High Level Expert Group on Scientific Data
A submission to the European Commission
October 2010

Riding the wave: How Europe can gain from the rising tide of scientific data.

Final report of the High Level Expert Group on Scientific Data. October 2010.

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

The Data Harvest:
How sharing research data can yield knowledge, jobs and growth
An RDA Europe Report
December 2014

The Data Harvest: How sharing research data can yield knowledge, jobs and growth.

An RDA Europe Report. December 2014

https://rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html

RDA RESEARCH DATA ALLIANCE

Research Data Sharing without barriers

https://www.rd-alliance.org/

HLEG report on European Open Science Cloud

(October 2016)

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

Emergence of Cognitive Technologies

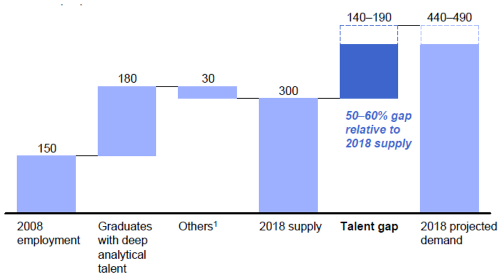
(IBM Watson and others)

The Fourth Paradigm of Scientific Research

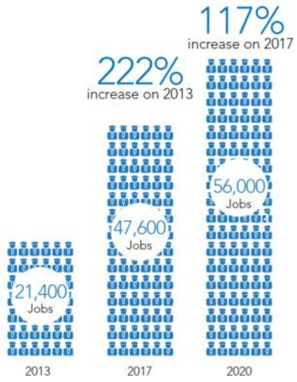
1. Theory, hypothesis and logical reasoning
2. Observation or Experiment
   - E.g. Newton observed apples falling to design his theory of mechanics
   - But Gallileo Galilei made experiments with falling objects from the Pisa leaning tower
3. Simulation of theory or model
   - Digital simulation can prove theory or model
4. Data-driven Scientific Discovery (aka Data Science)
   - More data beat hypnotized theory
   - e-Science as computing and Information Technologies empowered science
5. Human-computer driven science?

Demand for Data Science and data related professions

- McKinsey Global Institute on Big Data Jobs (2011)
  http://www.mckinsey.com/mgi/publications/big_data/index.asp
  - Estimated gap of 140,000 - 190,000 data analytics skills by 2018



- UK Big Data skills report 2014
  - 6400 UK organisations with 100+ staff will have implemented Big Data Analytics by 2020
  - Increase of Big Data jobs from 21,400 (2013) to 56,000 (2017)



- 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%
- HLEG report on European Open Science Cloud (October 2016) identified need for data experts and data stewards
  - Recommendation: Allocate 5% grant funding for Data management and preservation
  - Estimation: More than 80,000 data stewards (1 per every 20 scientists)
  - Core Data Experts (as defined) need to be trained and their career perspective improved

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
- Launch **Digital Skills and Jobs Coalition** (1st December 2016, Brussels) to develop comprehensive national digital skills strategies by mid-2017

Industry report on Data Science Analytics and Data enabled skills demand

- 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%
- PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017)
  - http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent
- Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017)
  - http://www.bhef.com/publications/quant-crunch-how-demand-data-science-skills-disrupting-job-market
  - https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF

OECD and UN on Digital Economy and Data Literacy

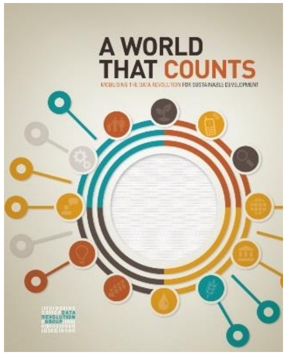
OECD

- Demand for new type of *“dynamic self-re-skilling workforce”*
- Continuous learning and professional development to become a shared responsibility of workers and organisations

[ref] SKILLS FOR A DIGITAL WORLD, OECD, 25-May-2016
http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/ICCP/IIS(2015)10/FINAL&docLanguage=En

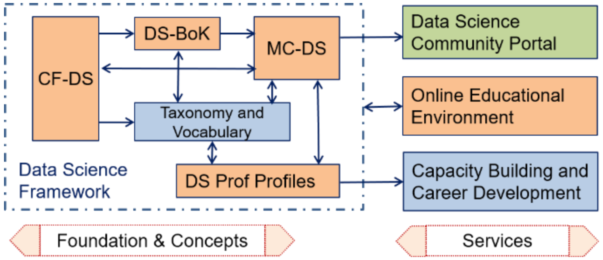
UN

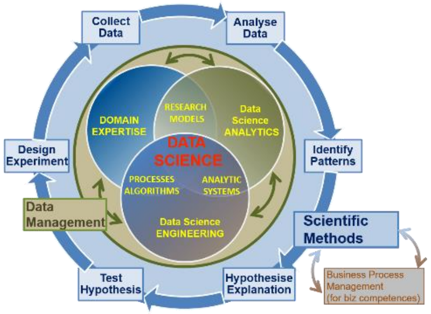
- Data Revolution Report "A WORLD THAT COUNTS" Presented to Secretary-General (2014)
  http://www.undatarevolution.org/report/
- Data Literacy is defined as key for digital revolution
- **Data literacy** = critically analyse data collected and data visualised

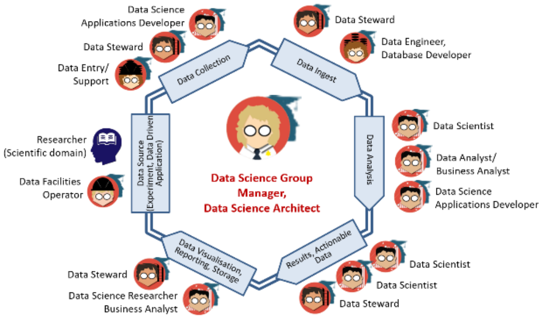


EDISON Services for Data Science and Skills Management and Tailored Education

- 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)

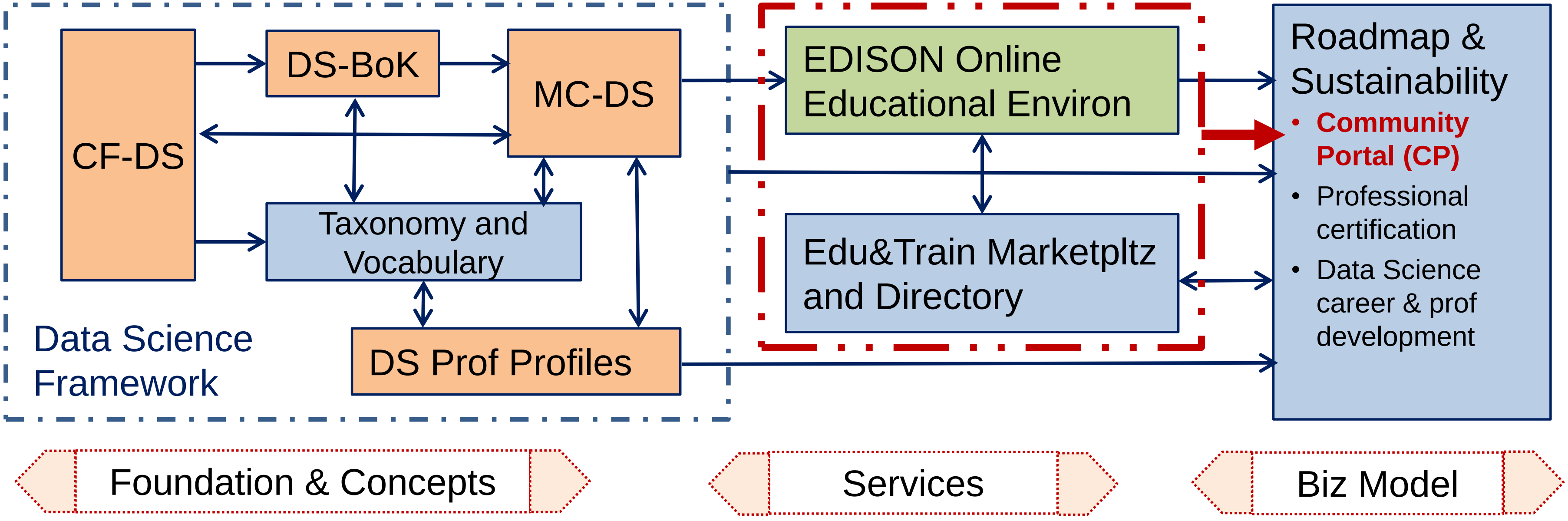






Data Science Applications Developer
Data Steward
Data Entry/ Support
Researcher (Scientific domain)
Data Facilities Operator
Data Steward
Data Science Researcher Business Analyst
Data Steward
Data Engineer, Database Developer
Data Scientist
Data Analyst/ Business Analyst
Data Science Applications Developer
Data Scientist
Data Scientist
Data Steward
Data Collection
Data Ingest
Data Analysis
Results, Actionable Data
Data Visualisation, Reporting, Storage
Data Source (Experiment, Data Driven Application)
Data Science Group Manager, Data Science Architect

EDISON Data Science Framework (EDSF)



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

Methodology

- ESDF development based on job market study, existing practices in academic, research and industry.
- Review and feedback from the ELG, expert community, domain experts.
- Input from the champion universities and community of practice.

What challenges with skills management EDSF can help to address?

1. Guide researchers in using right methods and tools, latest Data Analytics technologies to extracting value from scientific data
2. Educate and train RI engineers dev to build modern data intensive research infrastructure and understand trends and project for future
3. Develop new data analytics tools and ensure continuous improvement (agile model, DevOps)
4. Correctly organise and manage data, make them accessible (adhering FAIR principles), education new profession of Data Stewards
5. Help managers to facilitate career dev for researchers and organise effective teams
6. Ensure skills and expertise sustain in organisation
7. Help research institutions to sustain in competition with industry and business in data science talent hunting

EDSF: How CF-DS was constructed

- Background: Standards and Best Practices
- Jobs market analysis: Demanded Data Science Competences and Skills

Background: Standards and Best Practices

- e-CFv3.0 - European e-Competence Framework for IT
  - Structured by 4 Dimensions and organizational processes
    - Competence Areas: Plan – Build – Run – Enable - Manage
    - Competences: total defined 40 competences
    - Proficiency levels: identified 5 levels linked to professional education levels
    - Skills and Knowledge
- CWA 16458 (2012): European ICT Professional Profiles Family Tree
  - Defines 23 ICT profiles for common ICT jobs
- ESCO (European Skills, Competences, Qualifications and Occupations) framework
  - Standard for European job market since 2016
  - Expected inclusion of the Data Science occupations family – end 2017
- ACM Classification of Computer Science – CCS (2012)
- ACM Computer Science Body of Knowledge (CS-BoK) and ACM and IEEE Computer Science Curricula 2013 (CS2013)

Background: Standards and Best Practices

- e-CFv3.0 - European e-Competence Framew
  - Structured by 4 Dimensions and
    - Compete ... - Manage
    - 40 competences
    - ncy levels: identified 5 levels linked to professional education levels
    - Skills and Knowledge
- CWA 16458 (2012): European ICT Professional Profiles Family Tree
  - Defines 23 ICT profiles for common ICT jobs
- ESCO (European Skills, Competences, Qualifications and Occupations) framework
  - Standard for European job market since 2016
  - Expected inclusion of the Data Science occupations family – end 2017
- ACM Cla ... Science – CCS (2012)
- ACM Co ... CM and IEEE Computer Science Curricula 20

Currently work on e-CF4 is moved to CEN TC 428
To be extended with Data Science competences

New Joint Initiative ACM, IEEE, ASA, AAAS, AIS, ACH
To develop Data Science curriculum

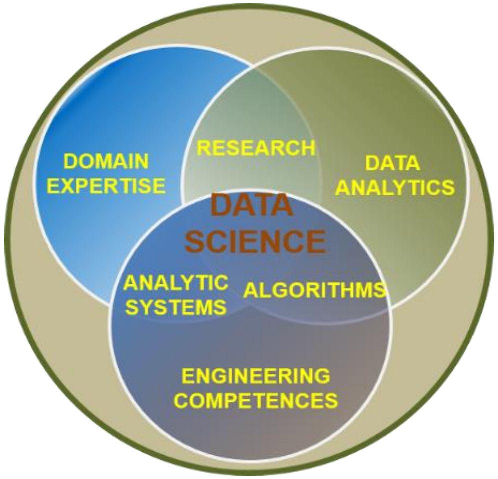
Jobs market analysis: Demanded Data Science Competences and Skills

- Initial Analysis (period Aug – Sept 2015) -> Continuous monitoring (in development)
  - IEEE Data Science Jobs (World but majority US)
    - Collected > 120, selected for analysis > 30
  - LinkedIn Data Science Jobs (NL)
    - Collected > 140, selected for analysis > 30
  - Existing studies and reports + numerous blogs & forums
- Analysis methods
  - Data analytics methods: classification, clustering, feature extraction
  - Research methods: Data collection - Hypothesis – Artefact - Evaluation
  - Expert evaluation by EDISON Liaison Groups (ELG), multiple workshops

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*



[ref] Legacy: NIST BDWG definition of Data Science

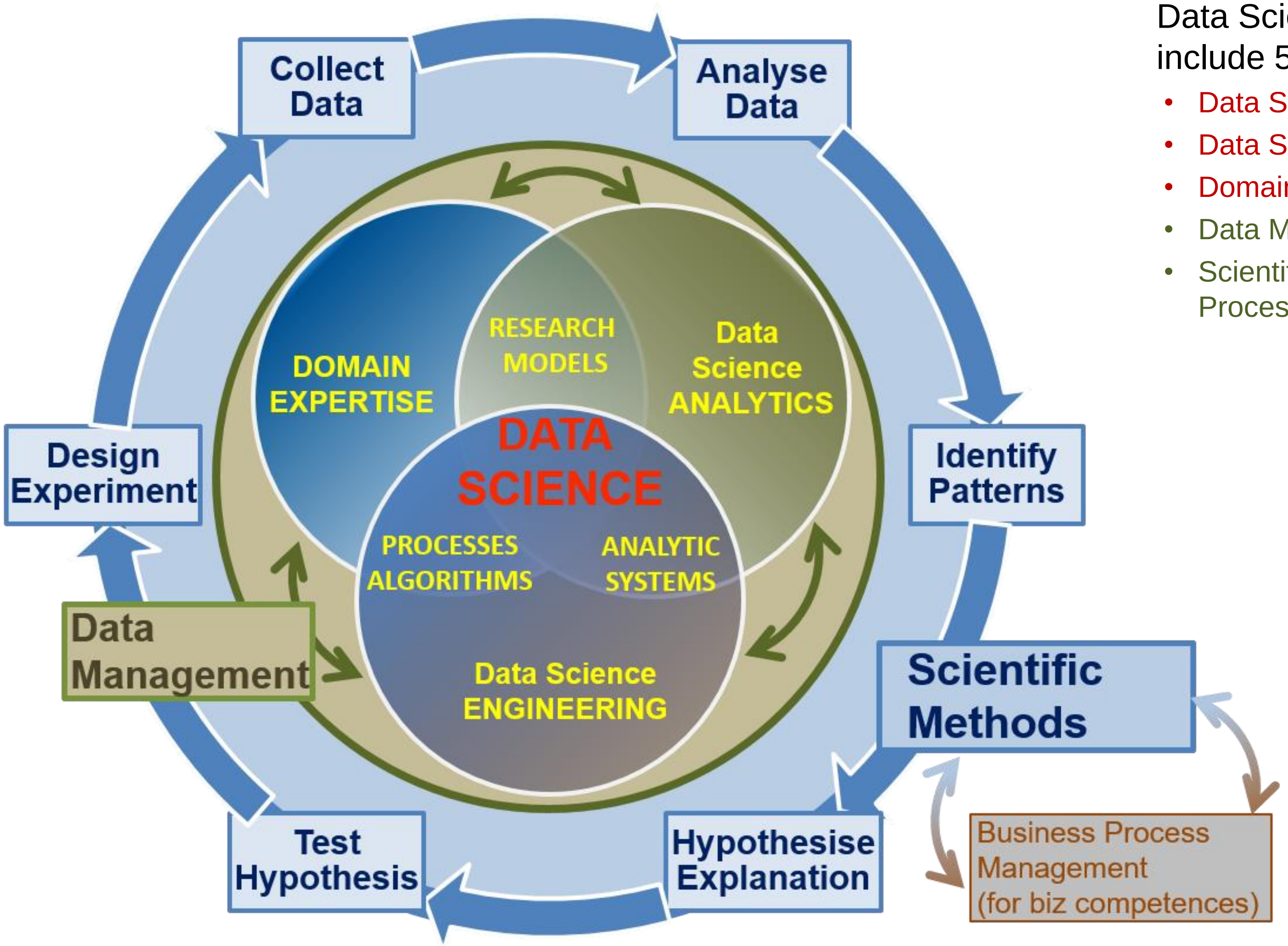
- *Other definitions to admit such features as*
  - Ability to solve variety of business problems
  - Optimize performance and suggest new services for the organisation
  - Develop a special mindset and be statistically minded, ***understand raw data** and **"appreciate data as a first class product"***

- ***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 Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Scientific Methods or Business Process Management

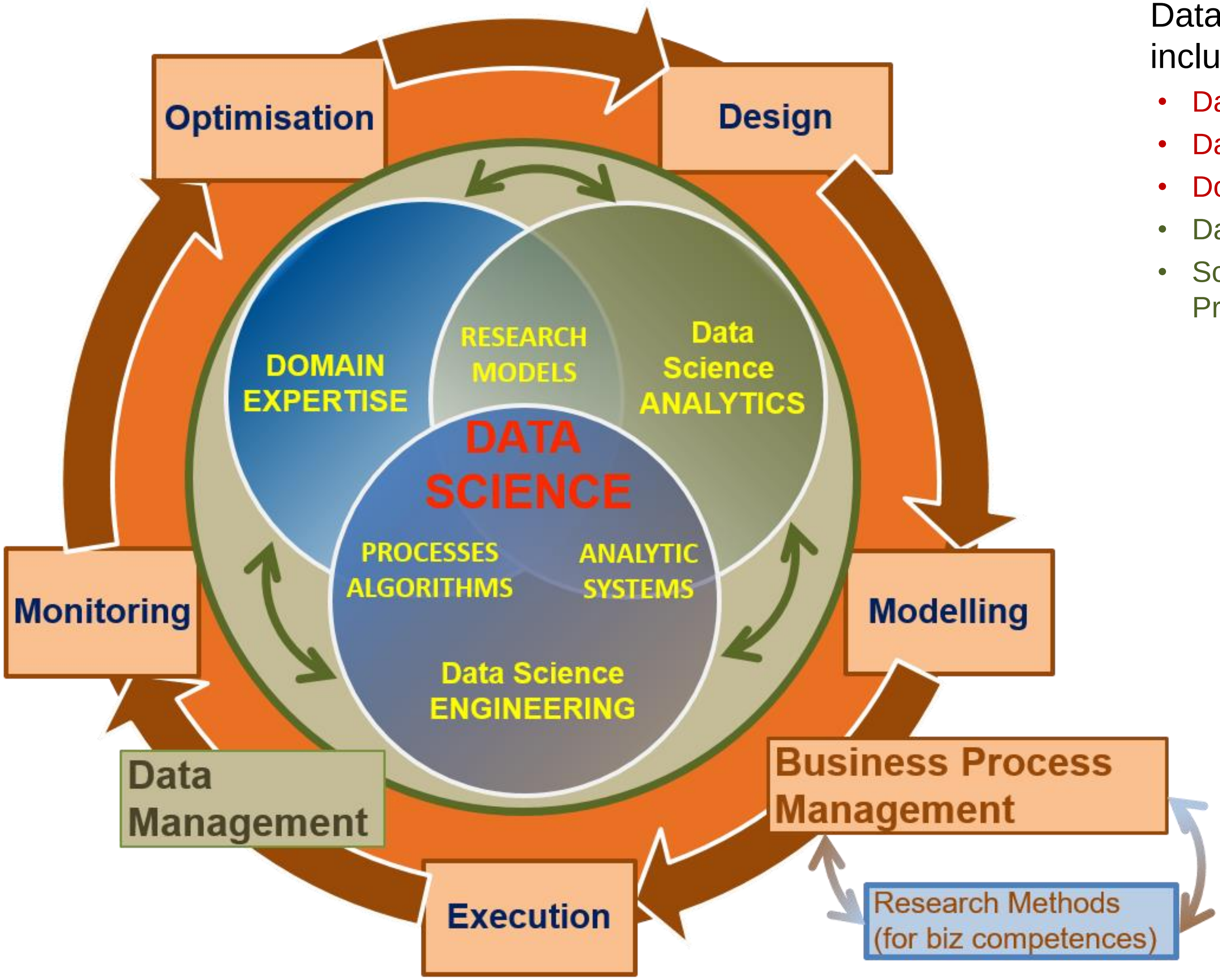
Scientific Methods

- 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 Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and 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

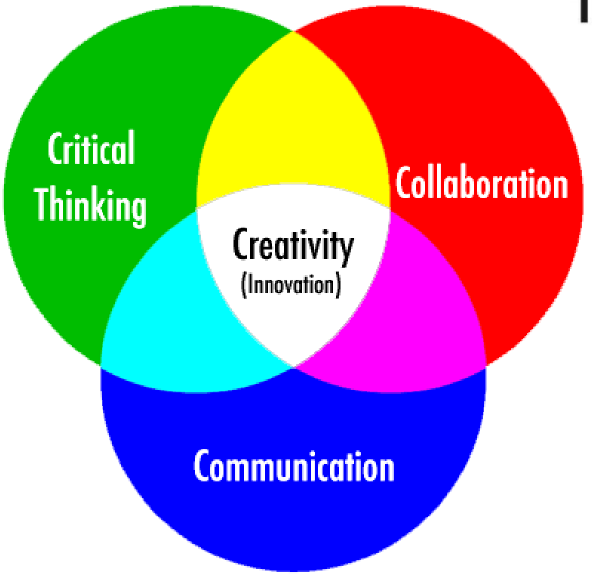
Key Data Science Analytics Competences by EDISON and DARE Project for APEC countries

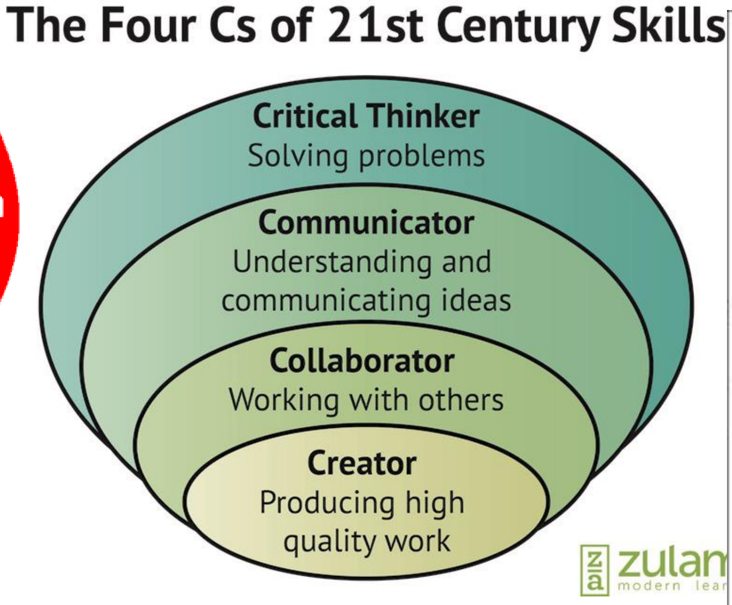
- Core/foundational competences (starting from entry level to expert level)
  - Statistics, Probability theory, mathematics, calculus
  - Statistical programming languages, frameworks, tools
  - Computational methods and document processing tools (including Excel, Office visualization, or similar)
  - Data Visualisation, and tools (e.g. Tableau, SPSS)
- Data Science Analytics (including Data Mining, Machine Learning)
  - Extended (data driven technologies): Optimization, simulation, etc.
- Data Science Engineering (including applications development, Big Data Infrastructure design and operation, Data Warehousing, Data and infrastructure Security)
- Research methods and Business process methods
- Domain related knowledge (e.g. scientific domains, business, industry, public sector)
- 21st Century Skills

21st Century Skills (DARE & BHEF & EDISON)

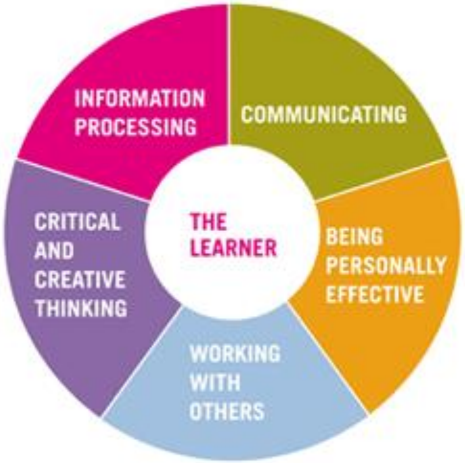
1. **Planning & Organizing**: Planning and prioritizing work to manage time effectively and accomplish assigned tasks
2. **Problem Solving**: Demonstrating the ability to apply critical thinking skills to solve problems by generating, evaluating, and implementing solutions
3. **Decision Making**: Applying critical thinking skills to solve problems encountered in the workplace
4. **Business Fundamentals**: Having fundamental knowledge of the organization and the industry
5. **Customer Focus**: Actively look for ways to identify market demands and meet customer or client needs
6. **Working with Tools & Technology**: Selecting, using, and maintaining tools and technology to facilitate work activity
7. **Dynamic (self-) re-skilling**: Continuously monitor individual knowledge and skills as shared responsibility between employer and employee

21st Century Skills – Different views of the same




The Four Cs of 21st Century Skills
Critical Thinker
Solving problems
Communicator
Understanding and communicating ideas
Collaborator
Working with others
Creator
Producing high quality work
zulan
modern lear

"I expect you all to be independent, innovative, critical thinkers who will do exactly as I say!"

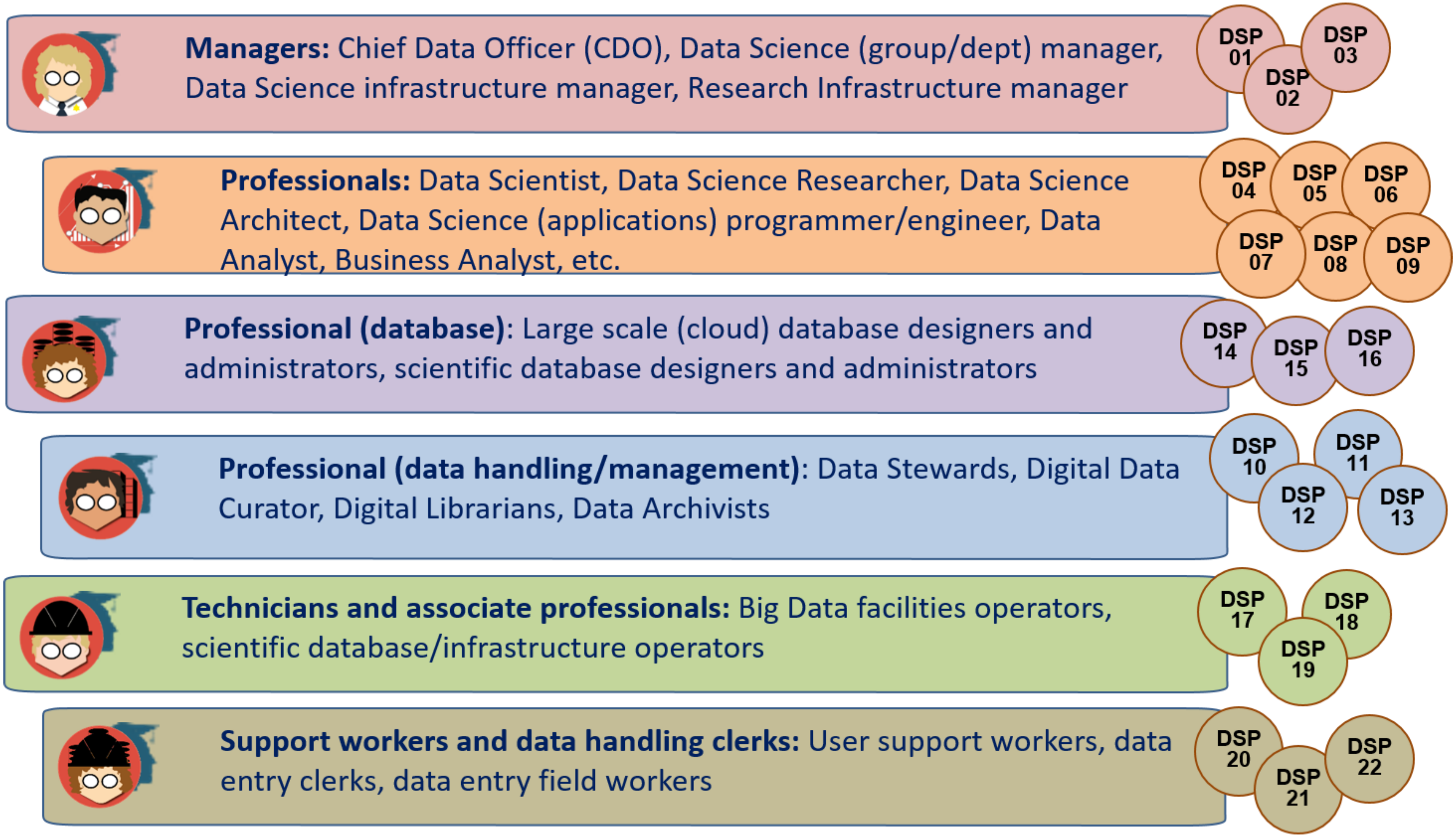


- New model in skills management: Shared responsibility between employee and employer
- Millennials factor: development and mobility

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

Data Science Professions Family



Managers: Chief Data Officer (CDO), Data Science (group/dept) manager, Data Science infrastructure manager, Research Infrastructure manager

DSP 01, DSP 02, DSP 03

Professionals: Data Scientist, Data Science Researcher, Data Science Architect, Data Science (applications) programmer/engineer, Data Analyst, Business Analyst, etc.

DSP 04, DSP 05, DSP 06, DSP 07, DSP 08, DSP 09

Professional (database): Large scale (cloud) database designers and administrators, scientific database designers and administrators

DSP 14, DSP 15, DSP 16

Professional (data handling/management): Data Stewards, Digital Data Curator, Digital Librarians, Data Archivists

DSP 10, DSP 11, DSP 12, DSP 13

Technicians and associate professionals: Big Data facilities operators, scientific database/infrastructure operators

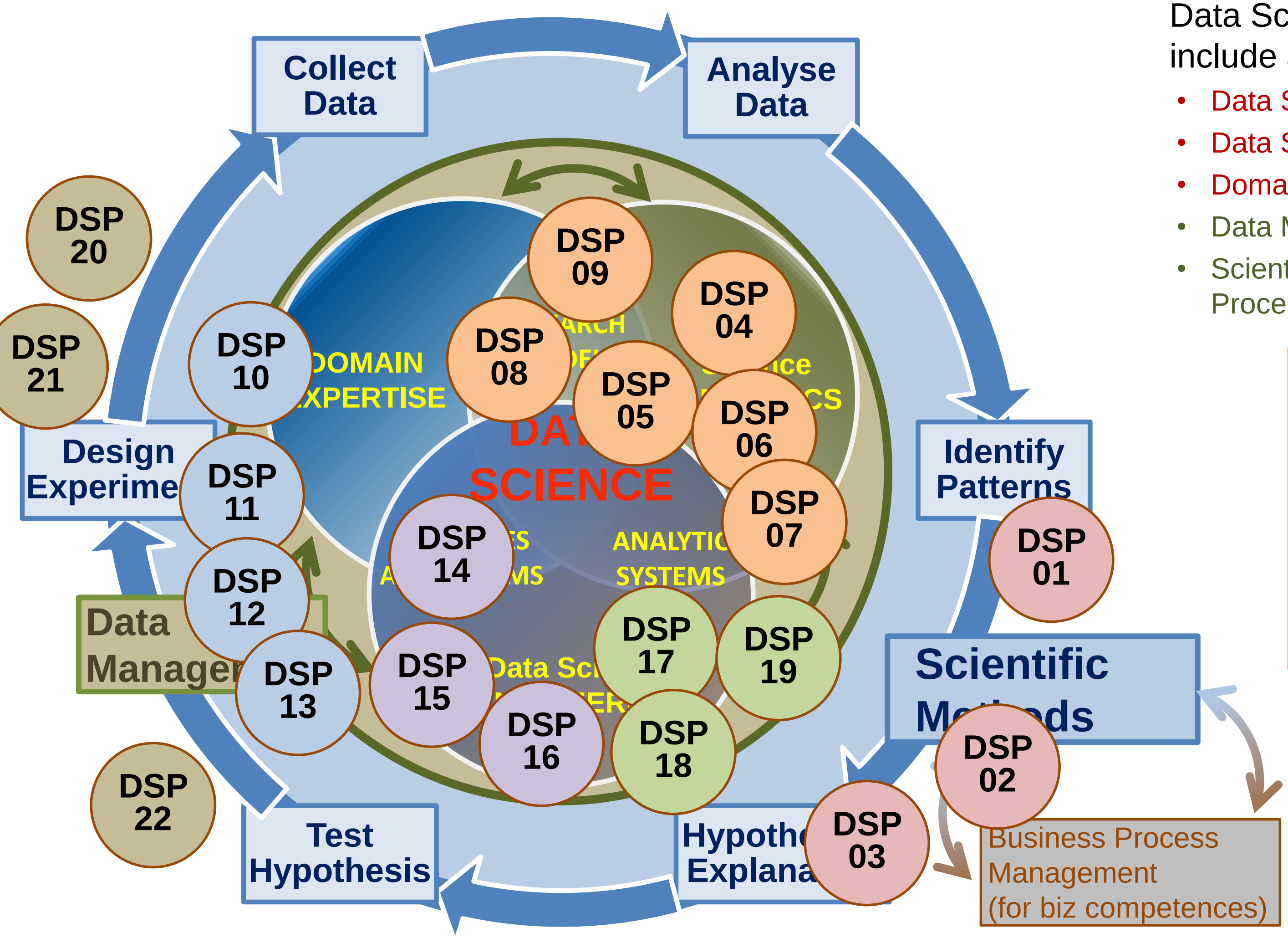
DSP 17, DSP 18, DSP 19

Support workers and data handling clerks: User support workers, data entry clerks, data entry field workers

DSP 20, DSP 21, DSP 22

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

Mapping DS Profiles to Competence Map



Data Science Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Scientific Methods or Business Process Management

Business Process Operations/Stages

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

Scientific Methods

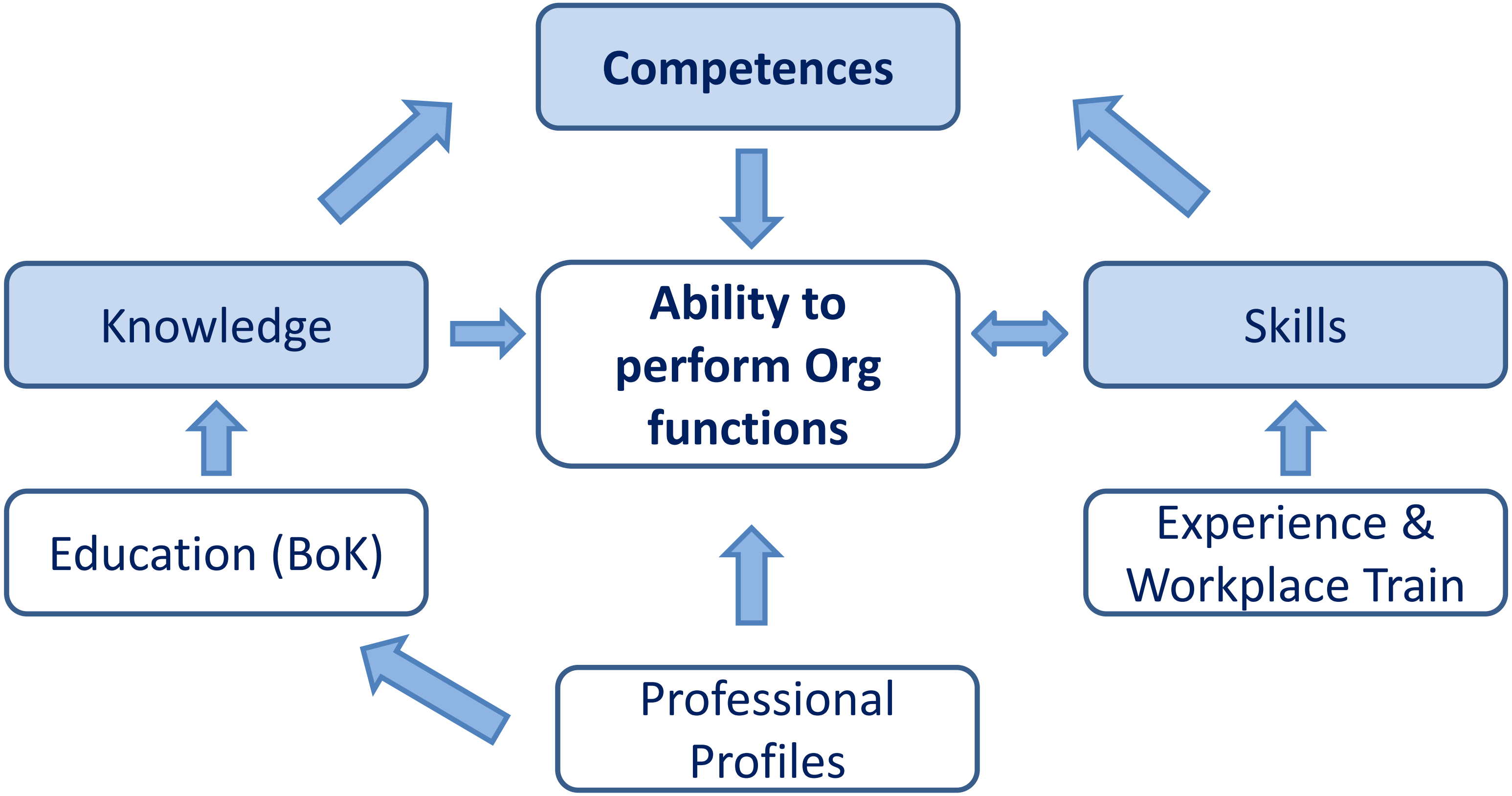
- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

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

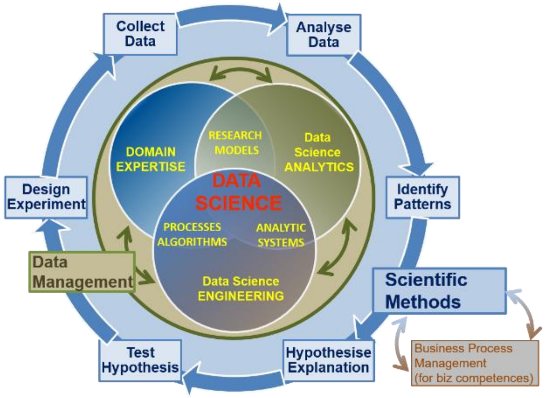
Competences Map to Knowledge and Skills

- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results



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

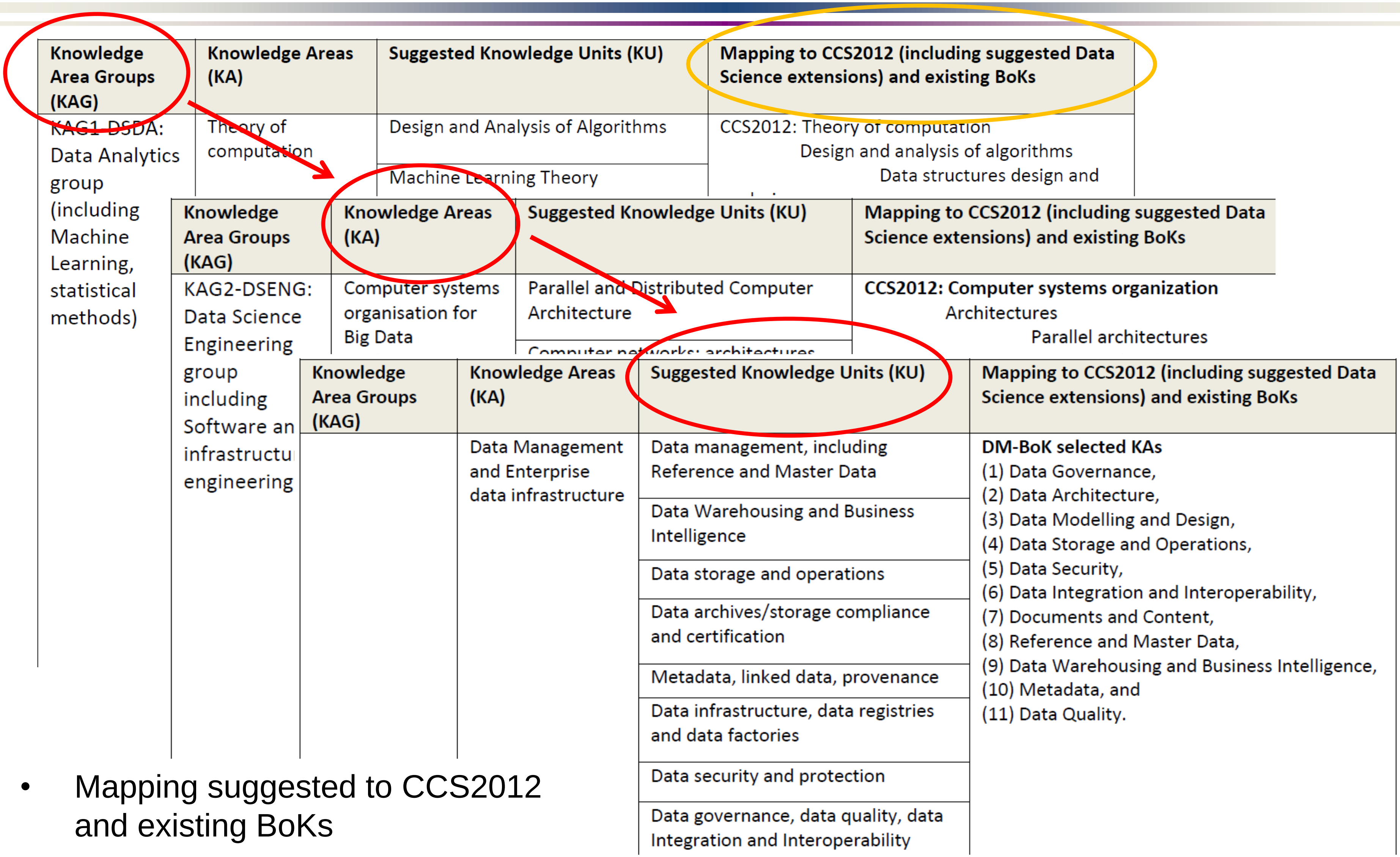
- Data Science domain knowledge to be defined by related expert groups

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 DS-BoK Knowledge Areas definition and mapping to existing BoKs and CCS (2012)



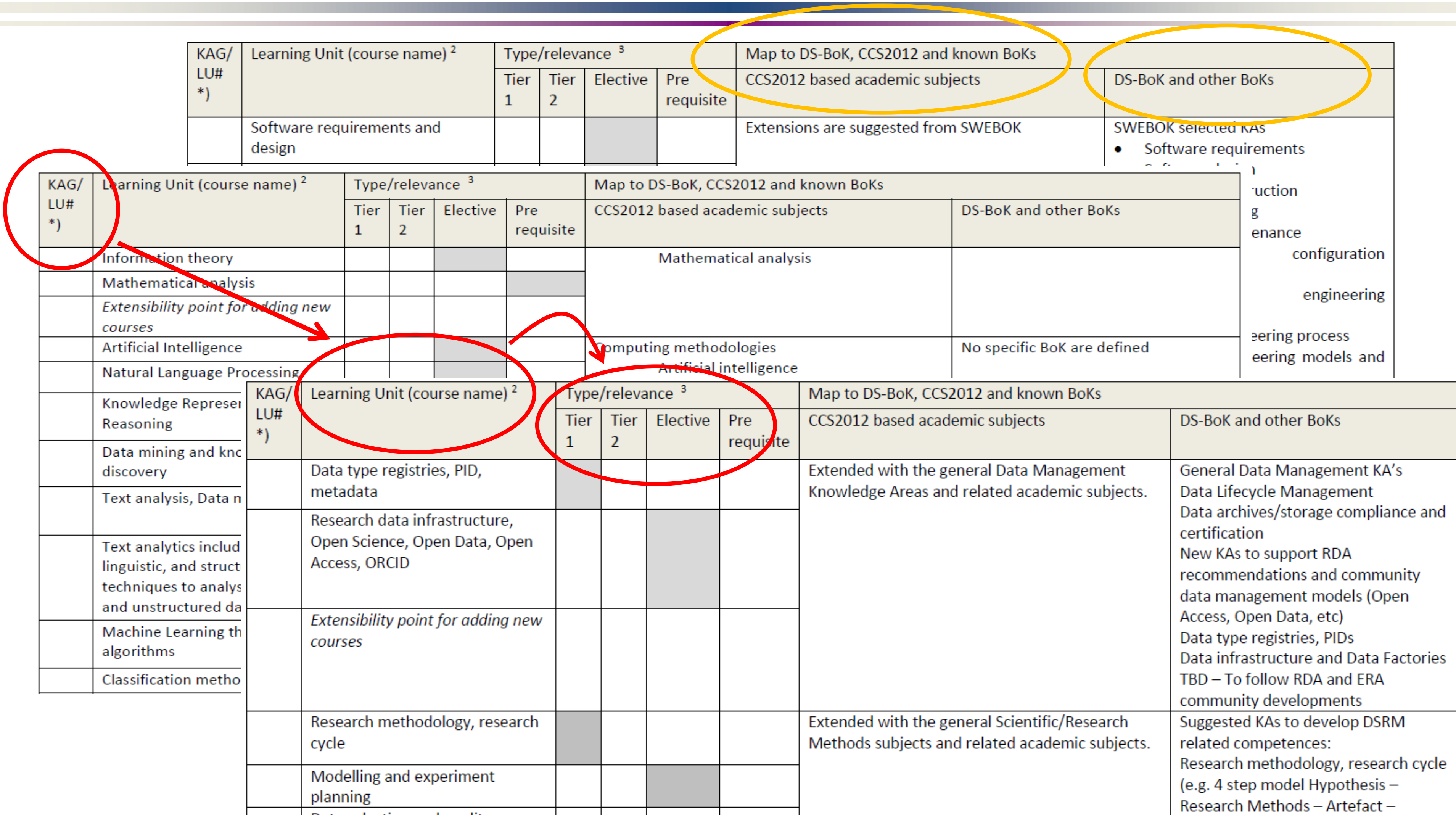
Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
KAG1-DSDA: Data Analytics group (including Machine Learning, statistical methods)	Theory of computation	Design and Analysis of Algorithms	CCS2012: Theory of computation Design and analysis of algorithms Data structures design and
		Machine Learning Theory	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
KAG2-DSENG: Data Science Engineering group including Software an infrastructu engineering	Computer systems organisation for Big Data	Parallel and Distributed Computer Architecture	**CCS2012: Computer systems organization** Architectures Parallel architectures
		Computer networks: architectures	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
	Data Management and Enterprise data infrastructure	Data management, including Reference and Master Data	**DM-BoK selected KAs** (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, and (11) Data Quality.
		Data Warehousing and Business Intelligence	
		Data storage and operations	
		Data archives/storage compliance and certification	
		Metadata, linked data, provenance	
		Data infrastructure, data registries and data factories	
		Data security and protection	
		Data governance, data quality, data Integration and Interoperability	

- Mapping suggested to CCS2012 and existing BoKs

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



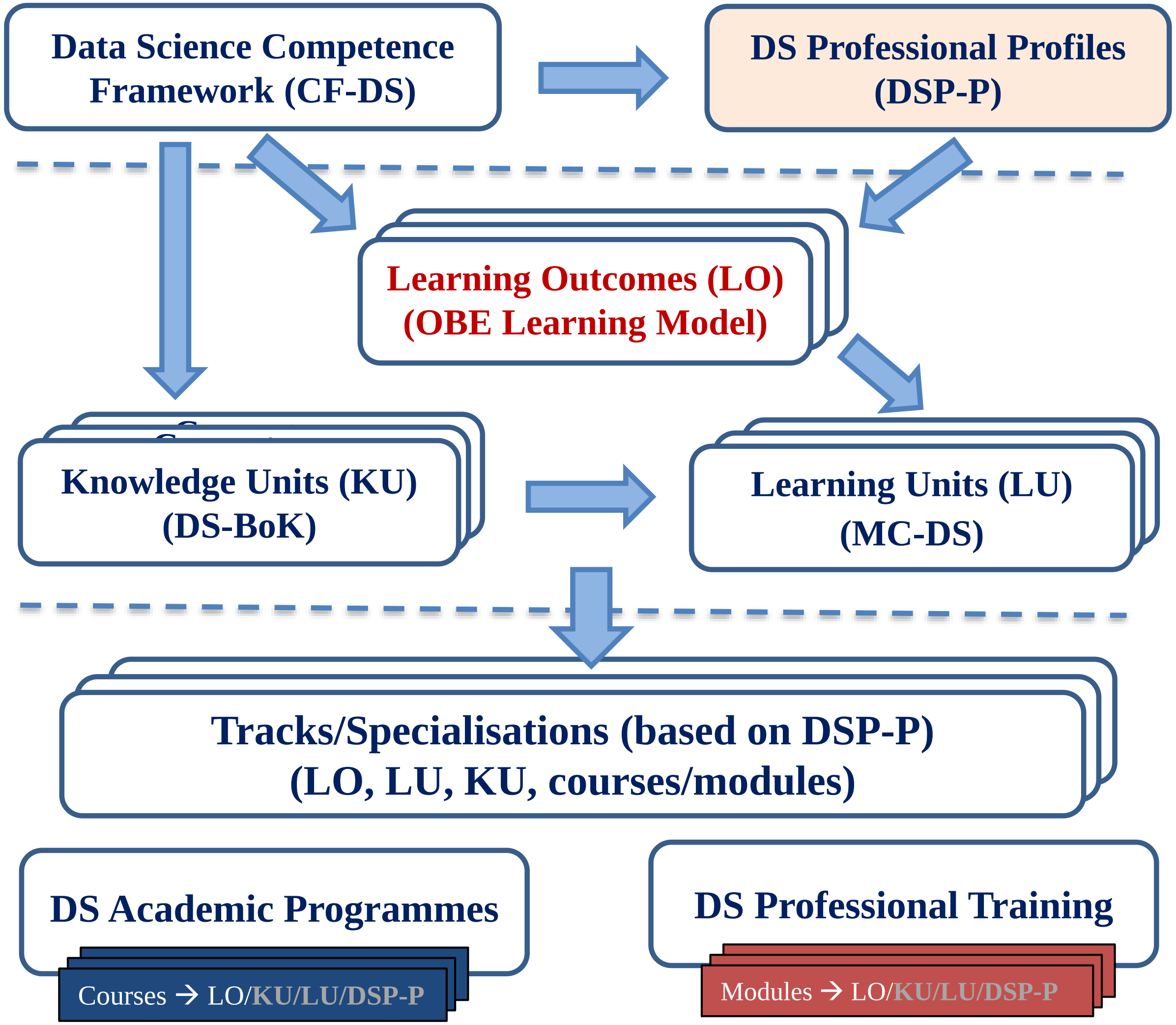
KAG/ LU# *)	Learning Unit (course name) [2]	Type/relevance [3]				Map to DS-BoK, CCS2012 and known BoKs	
		Tier 1	Tier 2	Elective	Pre requisite	CCS2012 based academic subjects	DS-BoK and other BoKs
	Software requirements and design			▨		Extensions are suggested from SWEBOK	SWEBOK selected KAs • Software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

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 Represen 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


Data Science Competence Framework (CF-DS)
DS Professional Profiles (DSP-P)
Learning Outcomes (LO) (OBE Learning Model)
Knowledge Units (KU) (DS-BoK)
Learning Units (LU) (MC-DS)
Tracks/Specialisations (based on DSP-P) (LO, LU, KU, courses/modules)
DS Academic Programmes
Courses → LO/KU/LU/DSP-P
DS Professional Training
Modules → LO/KU/LU/DSP-P

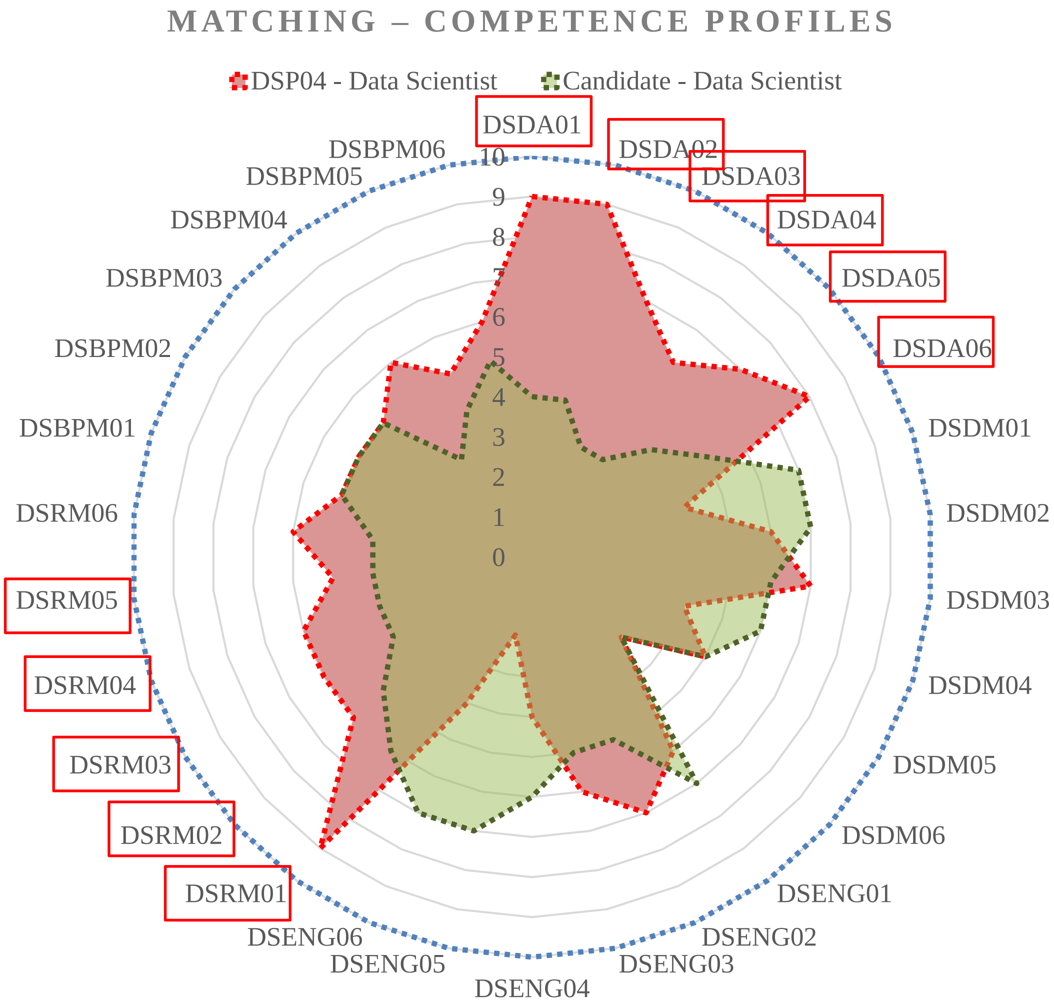
From Competences and DSP Profiles

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

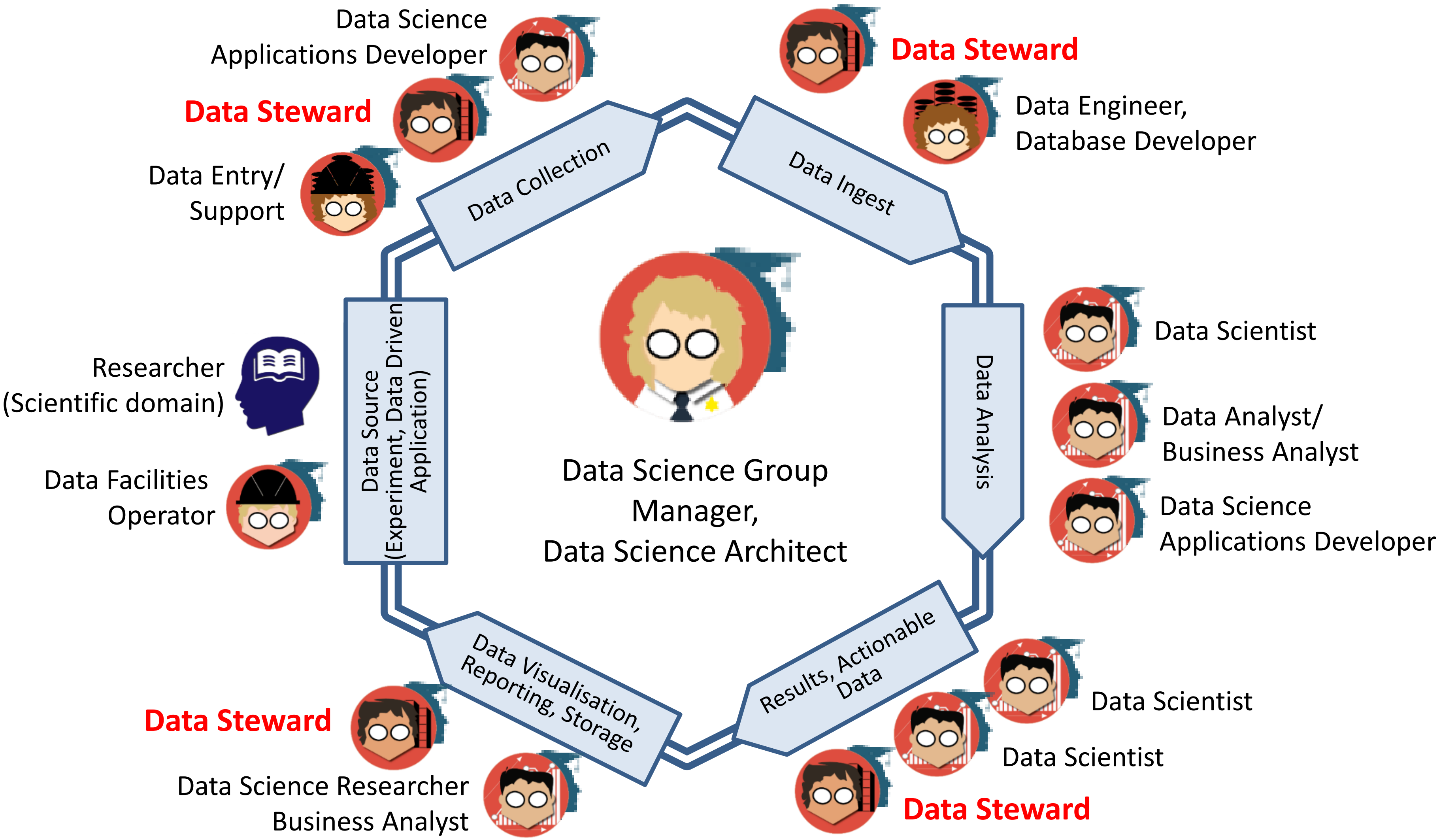


Individual Education/Training Path based on Competence benchmarking

- 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 Applications Developer
Data Steward
Data Entry/ Support
Data Steward
Data Engineer, Database Developer
Data Collection
Data Ingest
Researcher (Scientific domain)
Data Facilities Operator
Data Source (Experiment, Data Driven Application)
Data Science Group Manager, Data Science Architect
Data Analysis
Data Scientist
Data Analyst/ Business Analyst
Data Science Applications Developer
Data Visualisation, Reporting, Storage
Results, Actionable Data
Data Steward
Data Science Researcher Business Analyst
Data Scientist
Data Scientist
Data Steward

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)

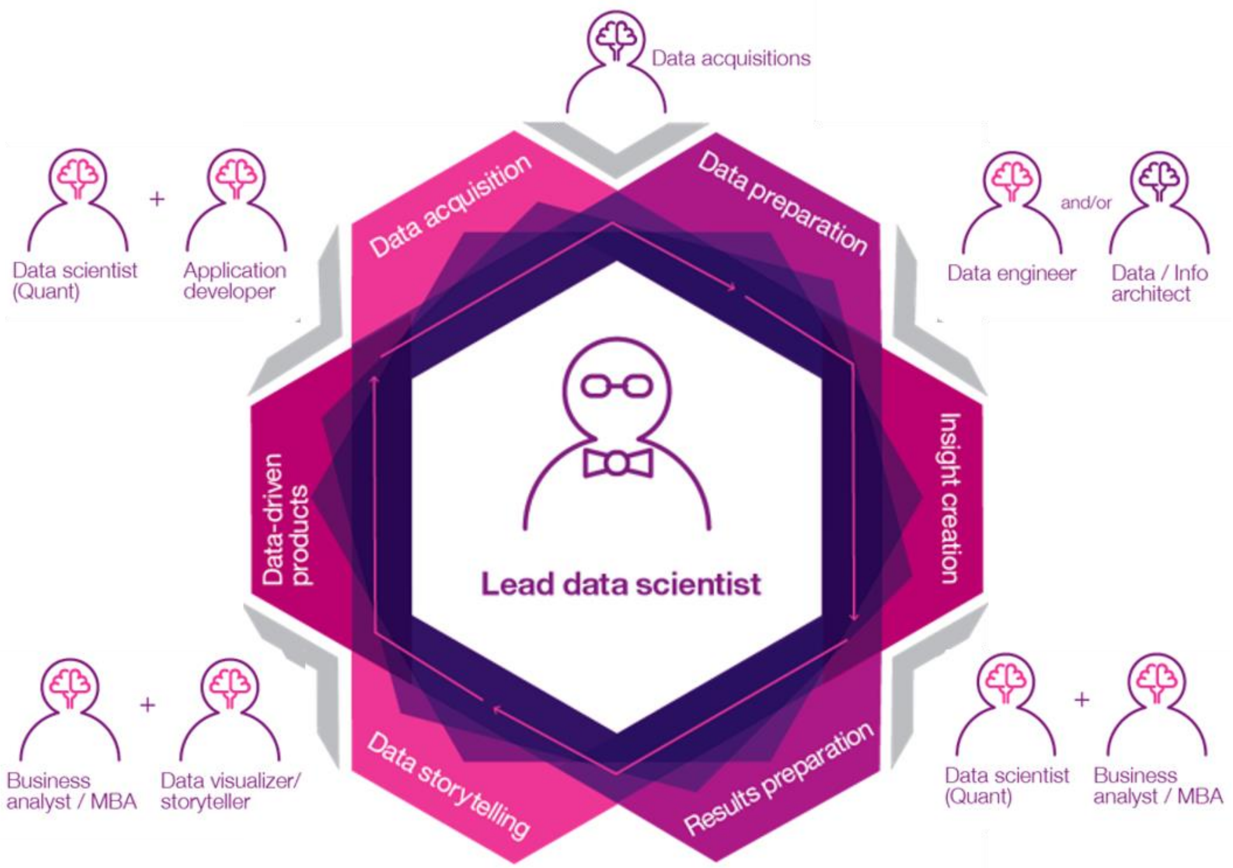
Estimated: Group of 10-12 data specialists for research institution of 200-300 research staff.

Growing role and demand for Data Stewards and data stewardship

Data Stewards – A rising new role in Data Science ecosystem

- Data Stewards as a key bridging role between Data Scientists as (hard)core data experts and scientific domain researchers
- Current definition of Data Steward (part of Data Science Professional profiles)
  - Data Steward is a data handling and management professional whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation. Data Steward creates data model for domain specific data, support and advice domain scientists/ researchers during the whole research cycle and data management lifecycle.

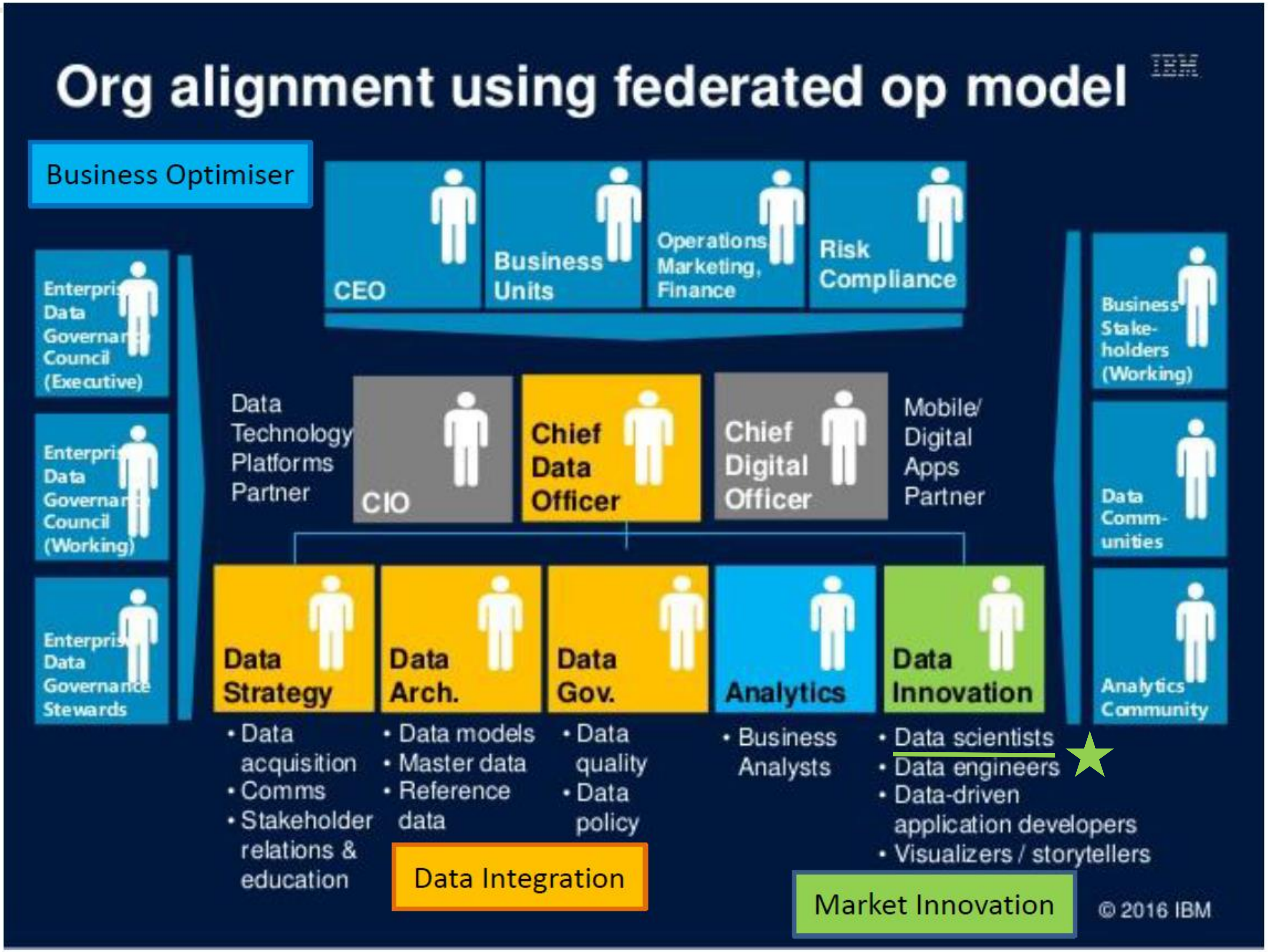
EXAMPLE: IBM emerging professions



[ref] Mastering the art of data science: How to craft cohesive teams that create business results, IBM Institute for Business Value, 2016

- Create data science teams with varied backgrounds and skills

[ref] Cortnie Abercrombie, What CEOs want from CDOs and how to deliver on it [online] http://www.slideshare.net/IBMBDA/what-ceos-want-from-cdos-and-how-to-deliver-on-it



Org alignment using federated op model

IBM

Business Optimiser

CEO

Business Units

Operations, Marketing, Finance

Risk Compliance

Enterprise Data Governance Council (Executive)

Enterprise Data Governance Council (Working)

Enterprise Data Governance Stewards

Data Technology Platforms Partner

CIO

Chief Data Officer

Chief Digital Officer

Mobile/ Digital Apps Partner

Business Stake-holders (Working)

Data Comm-unities

Analytics Community

Data Strategy
- Data acquisition
- Comms
- Stakeholder relations & education

Data Arch.
- Data models
- Master data
- Reference data

Data Gov.
- Data quality
- Data policy

Analytics
- Business Analysts

Data Innovation
- Data scientists
- Data engineers
- Data-driven application developers
- Visualizers / storytellers

Data Integration

Market Innovation

© 2016 IBM

New courses currently missing

- Data Management / Research Data Management
  - Data Curation, Data Stewardship
- Professional issues in Data Science
  - + Ethics and responsible use of Data Science

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)

Research Data Management Model Curriculum – Part of the EDISON Data Literacy Training

A. Use cases for data management and stewardship

- Preserving the Scientific Record

B. Data Management elements (organisational and individual)

- Goals and motivation for managing your data
- Data formats
- Creating documentation and metadata, metadata for discovery
- Using data portals and metadata registries
- Tracking Data Usage
- Handling sensitive data
- Backing up your data
- Data Management Plan (DMP) - to be a part of hands on session

Collaboration with the Research Data Alliance (RDA) on developing model curriculum on Research Data Literacy:

- Modular, Customisable, Localised, Open Access
- Supported by the network of trainers via resource swap board

C. Responsible Data Use Section (Citation, Copyright, Data Restrictions)

D. Open Science and Open Data (Definition, Standards, Open Data use and reuse, open government data)

- Research data and open access
- Repository and self- archiving services
- ORCID identifier for data
- Stakeholders and roles: engineer, librarian, researcher
- Open Data services: ORCID.org, Altmetric Doughnut, Zenodo

E. Hands on:

a) Data Management Plan design

b) Metadata and tools

c) Selection of licenses for open data and contents (e.g. Creative Common and Open Database)

Professional Issues in Data Science

- Data Science subjects/disciplines/components technologies
- Research Data Management and RDM Plan
  - Including data format, metadata
- Open Data and Open Science
- Data related skills and career management
  - Including Data Science certification
- Responsible Data Science and professional ethics
- 21st Century Skills
- Data protection, data privacy, data security

EDSF Recognition, Endorsement and Implementation

- **DARE (Data Analytics Rising Employment)** project by APEC (Asia Pacific Economic Cooperation)
  - DARE project Advisory Council meeting 4-5 May 2017, Singapore
- **PcW and BHEF Report "Investing in America's data science and analytics talent"** April 2017
  - Quotes EDSF and Amsterdam School of Data Science
- **Dutch Ministry of Education recommended EDSF** as a basis for university curricula on Data Science
  - Workshop "Be Prepared for Big Data in the Cloud: Dutch Initiatives for personalized medicine and health research & toward a national action programme for data science training", Amsterdam 28 June 2016
- **European Champion Universities network**
  - 1st Conference (13-14 July, UK), 2nd Conference (14-15 March, Madrid, Spain)
  - 3rd Conference 19-20 June 2017, Warsaw

Further developments and Next steps (1)

- Next EDSF release 2 (planned for June 2017) will link competences to skills and knowledge
- Final EDSF project deliverables (due August 2017) will include:
  - Data Science Education Sustainability Roadmap
    - Will involve wide consultation with experts community and also with EU policy makers
    - Will be reviewed by the EDISON Liaisons Groups (ELG)
  - Certification Framework for at least two levels of Data Science competences proficiency
    - Consultation with few certification providers is in the progress
- Toward EDSF and Data Science profession standardisation
  - ESCO (European Skills, Competences and Occupations) taxonomy – extending with the Data Science related occupations, competences and skills
  - CEN TC428 (European std body) – Extending current eCFv3.0 and ICT profiles towards e-CF4 with Data Science related competences
  - Work with the IEEE and ACM curriculum workshop to define Data Science Curriculum and extend current CCS2012 (Classification Computer Science 2012)
- Number of Case studies is planned in cooperation with active EU projects EDSA, EOSCpilot, BDVe, etc. (not limited to the project lifetime)

Further developments and Next steps (2)

- The EDISON project legacy will include
  (linked to the current project website and migrated to CP in the future)
  - EDSF – EDISON Data Science Framework
  - Data Science Community Portal (CP) - http://datasciencepro.eu/
  - EDISON project network including
    - EDISON Liaison Groups
    - Data Science Champions conference
    - Cooperative networks with European Research Infrastructures (e.g. HEP, Bioinformatics, Environment and Biodiversity, Maritime, etc),
    - International cooperative links BHEF, APEC, IEEE, ACM
- Applications and tools development
  - Prototypes will be produced in the timeline of the project but further development is a subject to additional funding
- Sustainability of the project legacy/products will be ensured by the project partners voluntarily for the period at least 3 yrs
  - EDSF will be maintained by UvA
  - CP by Engineering (Italy)

Further developments and Next steps (3)

- Further dissemination, engagement and outreach activity
  - Publishing final deliverables as BCP and books
  - Data Science Manifesto – Primarily focused on professional and ethical issues in Data Science, new type of professional
  - Inter-universities initiative “Data Science for UN’s Sustainable Development Goals” to focus in-curricula research (projects) on UN priority goals

- Wider engagement into EOSC activities related to RI Data related skills management and capacity building

European Open Science Cloud (EOSC)

HLEG report on European Open Science Cloud (October 2016) – Demand for Data Scientists/Stewards

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

https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

- Definition of the **Data Steward** as a distinctive role and profession
  - Core Data Experts need to be trained and their career perspective improved
- Estimation: More than 80,000 data stewards to serve 1.7 mln scientists in Europe (1 per every 20 scientists)
  - Based on 5% grant funding for Data management and preservation
- **Clash of cultures** between domain specialists and e-Infrastructure specialists (i.e. IT/Computer Science)

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)
- 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

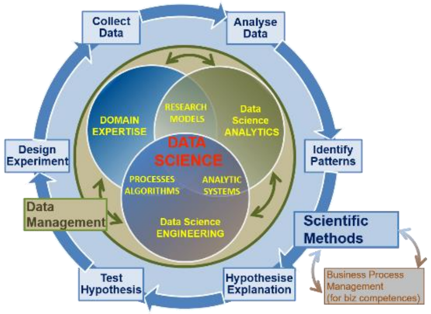
Summary: Services and References

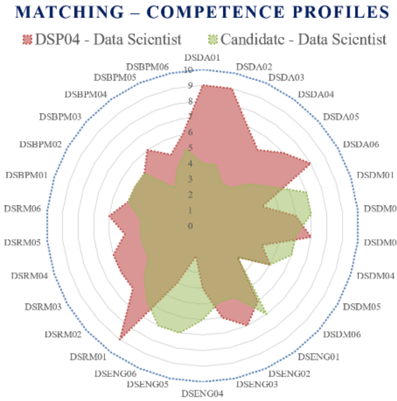
- EDISON Website http://edison-project.eu/
- EDISON Data Science Framework (EDSF) http://edison-project.eu/edison/edison-data-science-framework-edsf
- Directory of University programs http://edison-project.eu/university-programs-list
- Community Portal http://datasciencepro.eu/

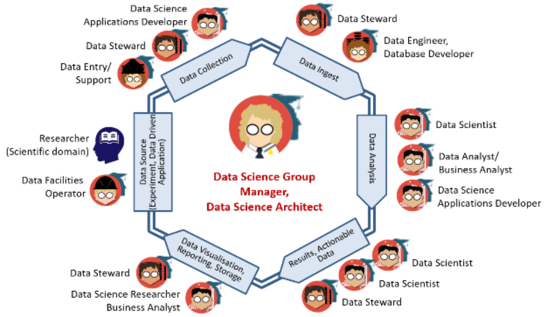
DATASCIENCEPRO

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

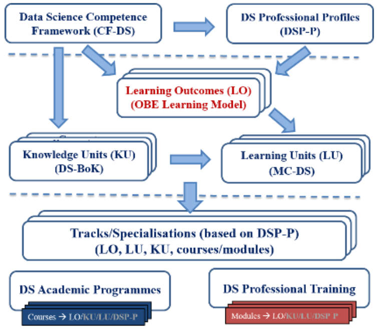
- Competences benchmarking and tailored training for practitioners
- Data Science Curriculum advice and design for universities
- Data Science team building and organizational roles profiling







Data Science Applications Developer
Data Steward
Data Entry/ Support
Researcher (Scientific domain)
Data Facilities Operator
Data Steward
Data Engineer, Database Developer
Data Scientist
Data Analyst/ Business Analyst
Data Science Applications Developer
Data Scientist
Data Scientist
Data Steward
Data Steward
Data Science Researcher Business Analyst
Data Collection
Data Ingest
Data Analysis
Results, Actionable Data
Data Visualisation, Reporting, Storage
Data Source (Experiment, Data Driven Application)
Data Science Group Manager, Data Science Architect



Data Science Competence Framework (CF-DS)
DS Professional Profiles (DSP-P)
Learning Outcomes (LO) (OBE Learning Model)
Knowledge Units (KU) (DS-BoK)
Learning Units (LU) (MC-DS)
Tracks/Specialisations (based on DSP-P) (LO, LU, KU, courses/modules)
DS Academic Programmes
Courses → LO/KU/LU/DSP-P
DS Professional Training
Modules → LO/KU/LU/DSP P

Discussion

- Questions
- Reflections
- Suggestions

Links to EDISON Resources

- 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

Other related links

- Amsterdam School of Data Science
  - https://www.schoolofdatascience.amsterdam/
  - https://www.schoolofdatascience.amsterdam/education/

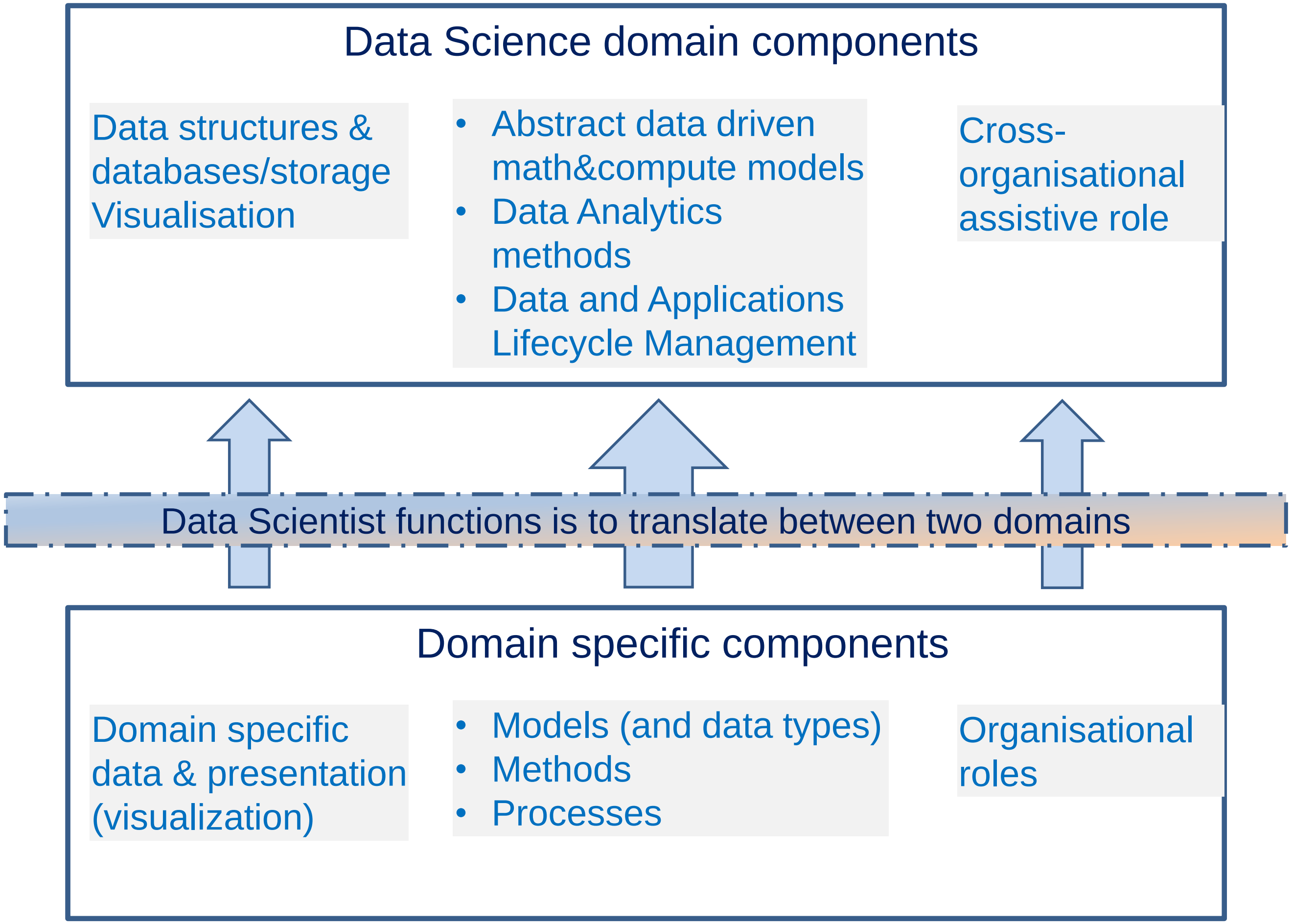
- Research Data Alliance interest Group on Education and Training on Handling of Research Data (IG-ETHRD)
  - https://www.rd-alliance.org/groups/education-and-training-handling-research-data.html
- PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017)
  - http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent
- Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017)
  - http://www.bhef.com/publications/quant-crunch-how-demand-data-science-skills-disrupting-job-market
  - https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF

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 Science domain components

Data structures & databases/storage Visualisation

- Abstract data driven math&compute models
- Data Analytics methods
- Data and Applications Lifecycle Management

Cross-organisational assistive role

Data Scientist functions is to translate between two domains

Domain specific components

Domain specific data & presentation (visualization)

- Models (and data types)
- Methods
- Processes

Organisational roles

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

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

EDISON Network and Engagement Activity (2)

- Workshops to promote a common approach towards addressing growing demand for Data Science and critical data competences and skills as required by **European Research Infrastructures (RI)**, future **European Open Science Cloud (EOSC)** and generally European **Digital Single Market (DSM)**.
  - **Joint EDISON and EC workshop** “Data Infrastructure Competences and Skills Framework: a European and Global Challenge” (Brussels, 9th February, 2016)
  - Joint IEEE, STC CC and RDA Workshop on Curricula and Teaching Methods (DTW2015 and DTW2016 collocated with IEEE CloudCom)
- EDISON initiated a set of **national action meetings** to address Data Science and digital skills by bringing together key stakeholders from universities, employer associations, and government
  - A first workshop jointly organised by the EDISON project and Dutch Ministry of Education, Culture and Science in June 2016 (during Netherlands Presidency in EU)

EDISON Network and Engagement Activity (3)

- European and international standardisation bodies and professional organisations
  - CEN TC426 Committee (former e-Competence Framework e-CFv3.0 workshop)
  - ESCO (European Skills, Competence, Occupations)
  - CEPIS and association **ICT Professionalism Europe** (co-signed 21 Nov 2016, Amsterdam)
- EDISON Booth at the Launch event of the **Digital Skills and Jobs Coalition: Boosting Europe's Digital skills**, 1 December 2016, Brussels
  - Part of actions toward European Digital Single Market (DSM) and
- Contribution to International standardization bodies, professional organisations and initiatives
  - **Business Higher Education Forum (BHEF) in USA**
  - **DARE project for APEC** (Asia Pacific Economic Cooperation) to develop a Data Analytics checklist for APEC countries
  - **Data Science Curriculum** Meeting of Professional and Academic Societies in USA (4 March 2017, Alexandria) including ACM