EDISON Data Science Framework (EDSF)

as a foundation for the Data Science education, training and sustainable skills development

Yuri Demchenko, EDISON Project
University of Amsterdam

e-IRG Workshop

8-9 June 2017, Malta

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

- Background
  - Recent EU Initiatives, European Digital Single Market (DSM) and demand for data enabled skills
- 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
- Use of EDSF for Data Science curricula design and organisational skills management
- Summary and discussion

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

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

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

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

HLEG report on European Open Science Cloud (October 2016) identified need for data experts and data stewards

- Estimation: More than 80,000 data stewards (1 per every 20 scientists)
- **Core Data Experts** need to be trained and their career perspective improved

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
  - 2.5 mln postings, 23% Data Scientist, 67% DSA enabled jobs
  - DSA enabled jobs growing at higher rate than main Data Science jobs
- Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017)
  - https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF
  - DSA enabled jobs takes 45-58 days to fill: 5 days longer than average

ICT and Data related Skills shortage (in life)

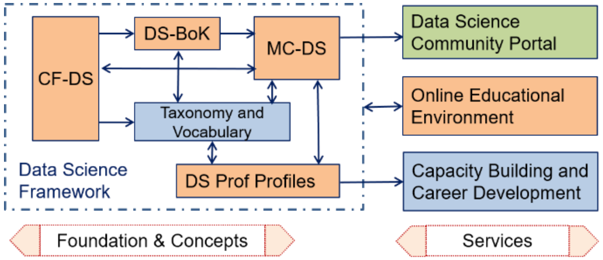
- Problems with hiring (skilled) ICT related staff
  - At least one year for training and acquiring experience
  - As soon as new employees are confident with their skills, they leave for big companies or industry
- Open Data Science/Stewards positions stay unfilled
  - In research institutions for months and years
  - In industry for months
- Companies/organisations want experienced Data Science workers
  - There is no time to acquire necessary experience
- Millennials factor
  - Do we understand difference of the millennials workforce?
- Challenge: How to obtain, train and sustain new digital (ICT and Data related) skills in organisations

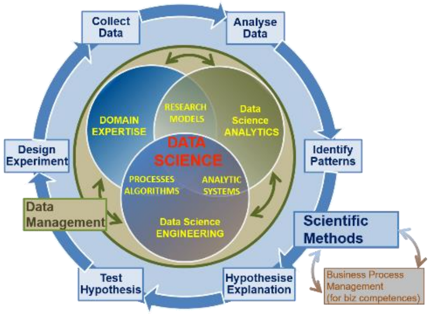
Sustainable ICT and Data Skills Development

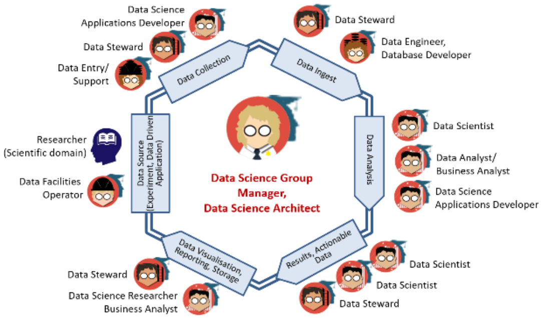
- HLEG report on EOSC rose question about critical need for Core Data Experts
  - Any changes since report publication in October 2016?
- Educate vs Train
  - Training is a short term solution
  - Education is a basis for sustainable skills development
- Technology focus changes every 3-4 years
  - Study: 50% of academic curricula are outdated at the time of graduation
- Lack of necessary skills leads to underperforming projects and organisations and loose of competitiveness
- Need to change skills management paradigm
  - **Dynamic (self-) re-skilling:** Continuous professional development and shared responsibility between employer and employee
  - Skills and career management as a part of professional orientation

EDISON Products for Data Science Skills Management and Tailored Education

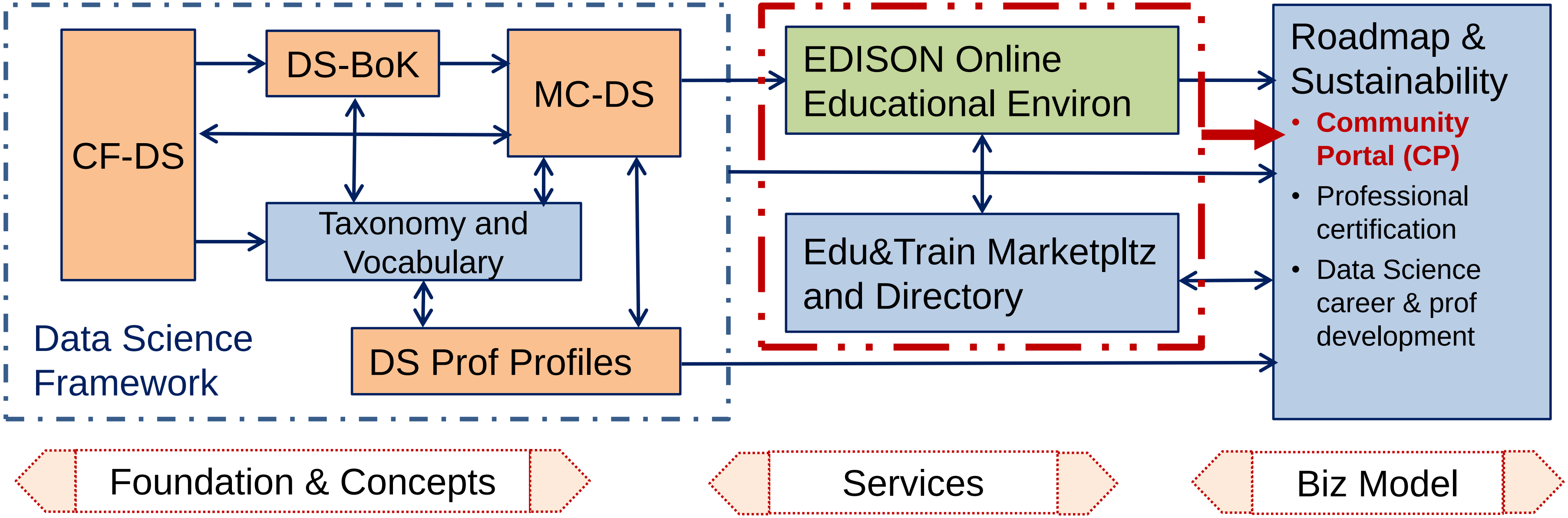
- EDISON Data Science Framework (EDSF)
  - – Compliant with EU standards on competences and professional occupations e-CFv3.0, ESCO
  - – Customisable courses design for targeted education and training
- Skills development and career management for Core Data Experts and related data handling professions
- 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 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

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

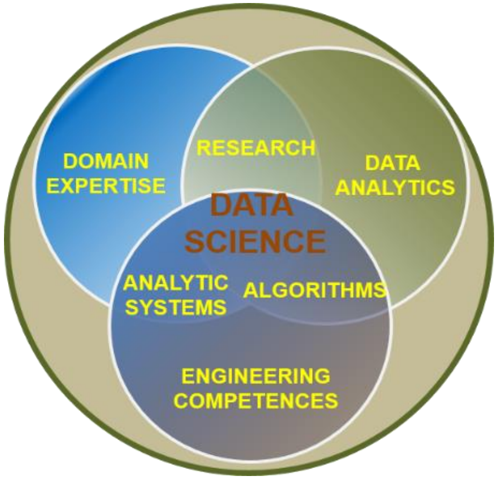
Methodology

- 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.

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*

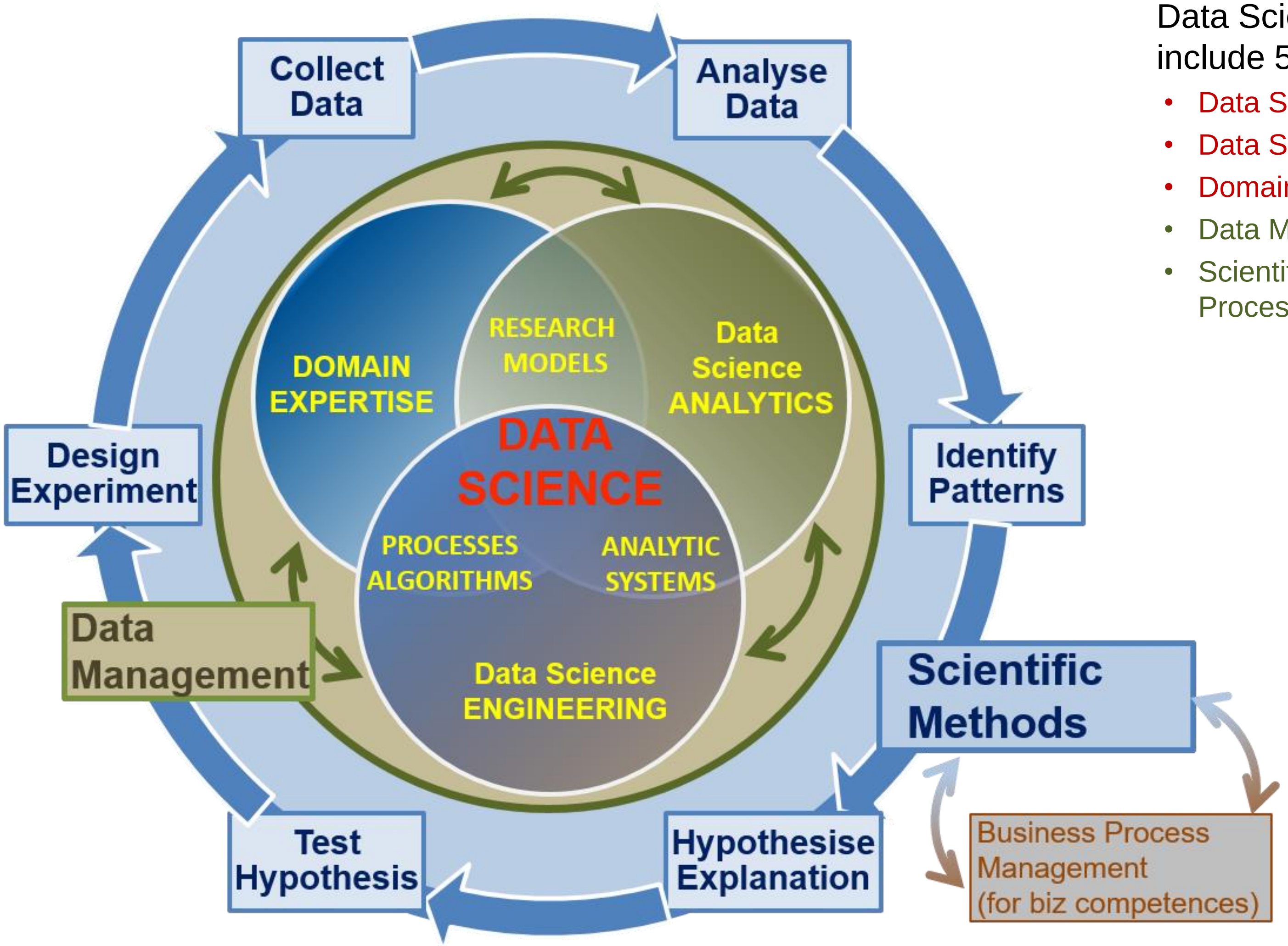


[ref] Legacy: NIST BDWG definition of Data Science

- *Profession is defined via **Competences** mapped to*
  - ***Skills and Knowledge***
  - ***Proficiency levels***

- ***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*

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

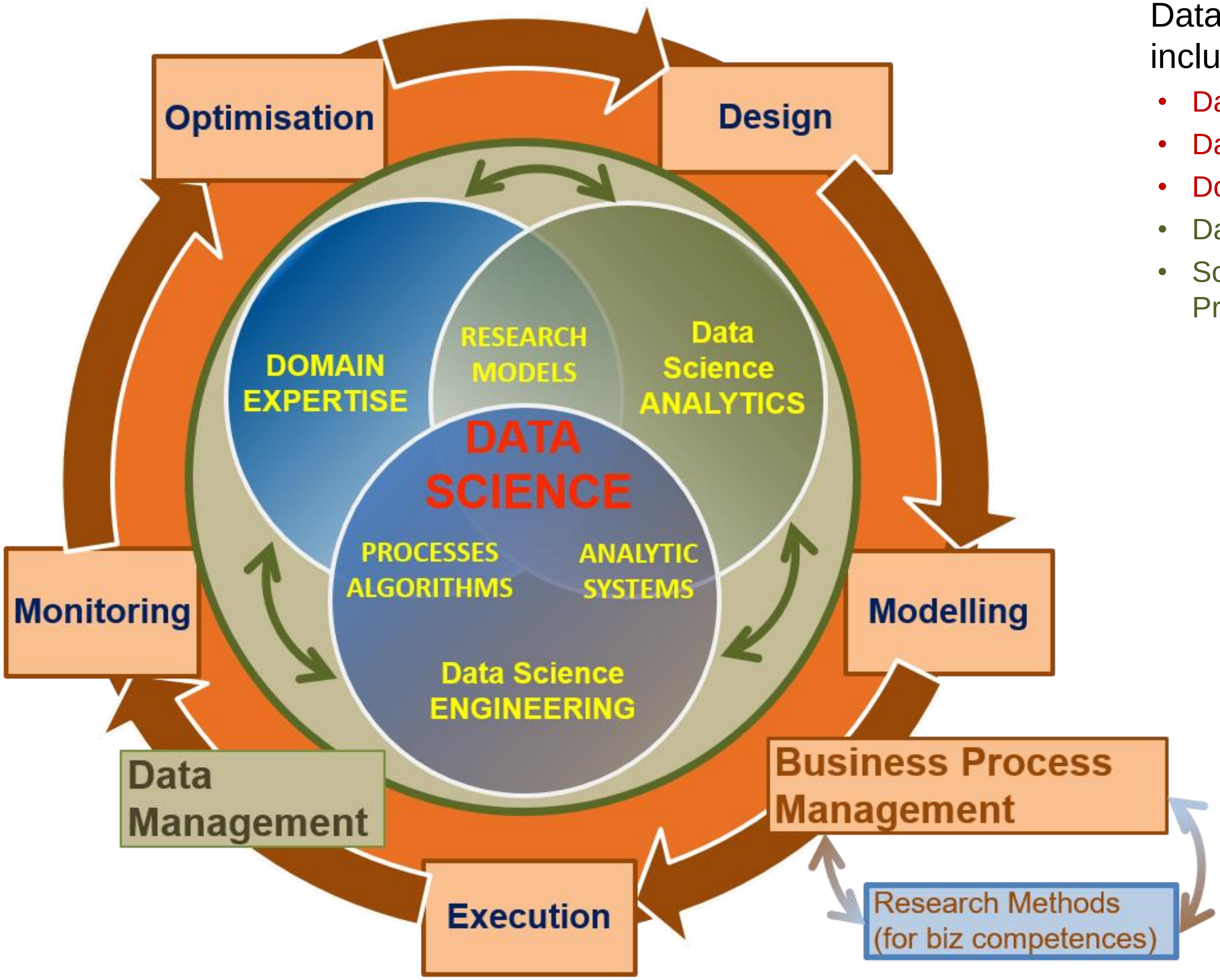
Scientific Methods

- 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

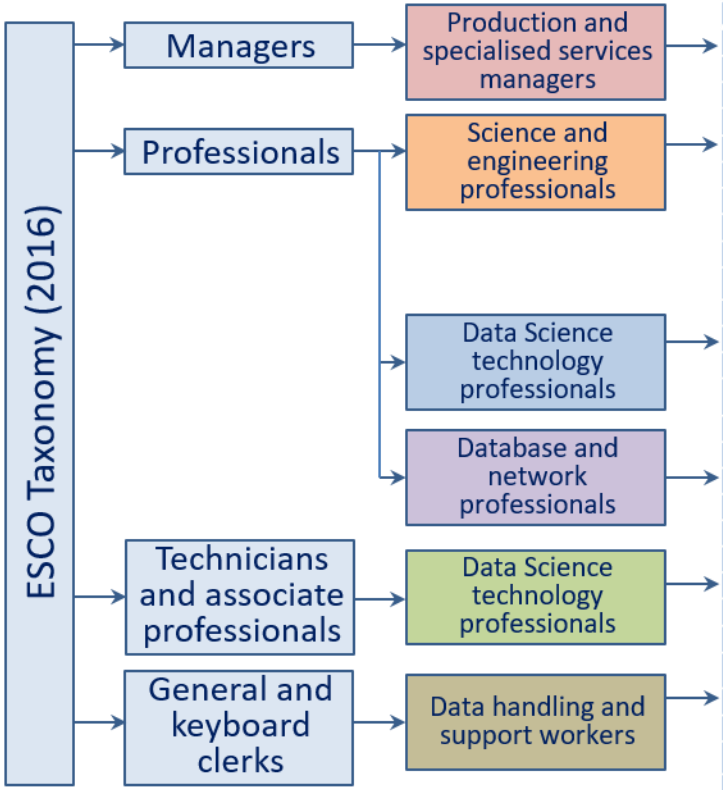
Identified Data Science *Skills/Experience* Groups

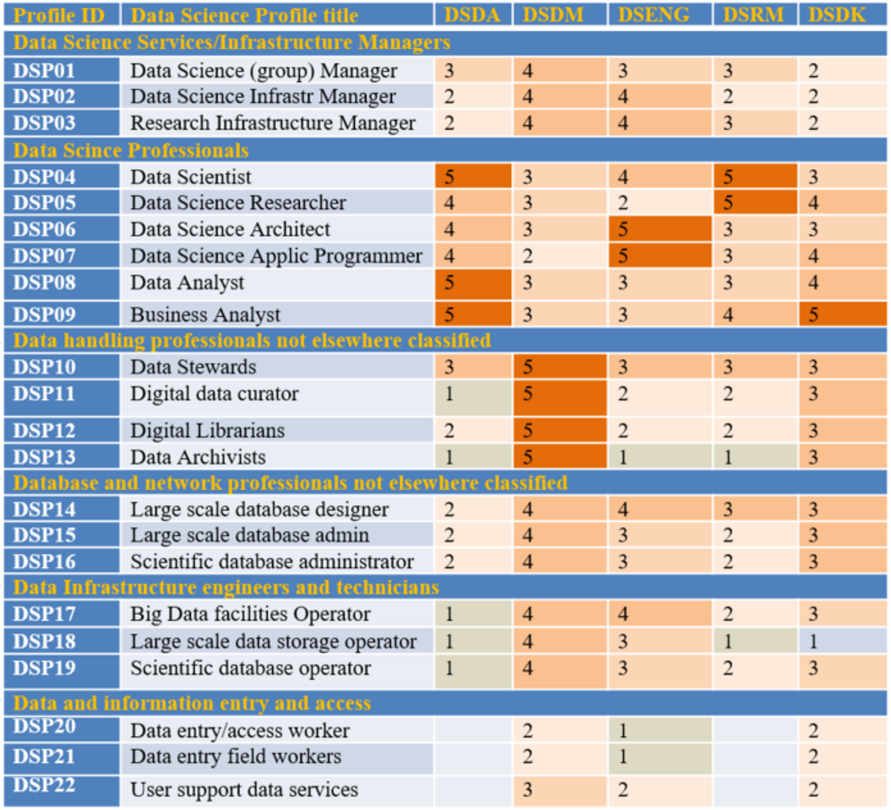
- **Group 1: Skills/experience related to competences**
  - Data Analytics and Machine Learning
  - Data Management/Curation (including both general data management and scientific data management)
  - Data Science Engineering (hardware and software) skills
  - Scientific/Research Methods or Business 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 21st Century Skills**
  - 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

DSP Profiles mapping to ESCO Taxonomy High Level Groups

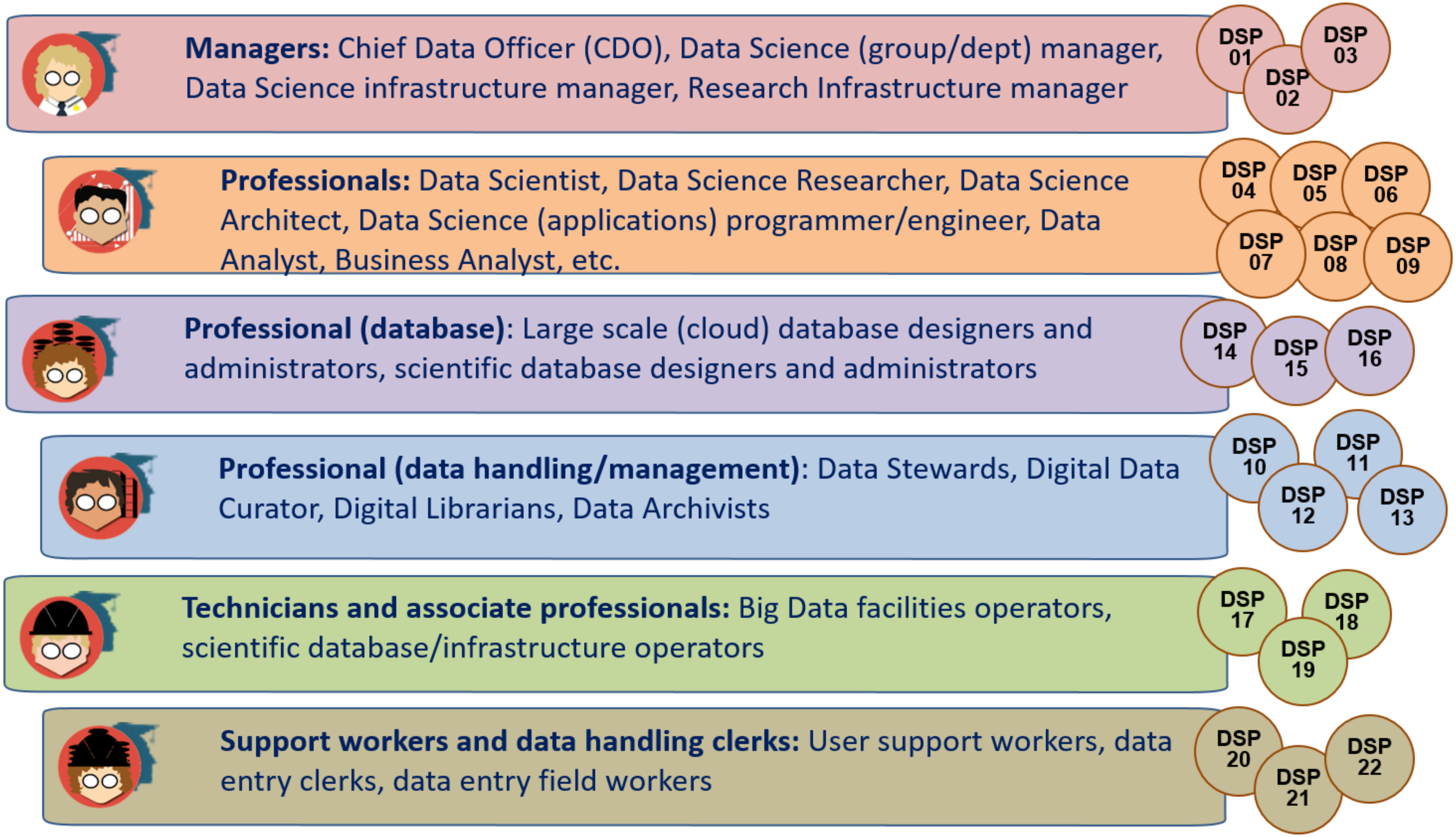

ESCO Taxonomy (2016)
Managers
Professionals
Technicians and associate professionals
General and keyboard clerks
Production and specialised services managers
Science and engineering professionals
Data Science technology professionals
Database and network professionals
Data Science technology professionals
Data handling and support workers



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 Science Professions Family



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

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

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

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

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

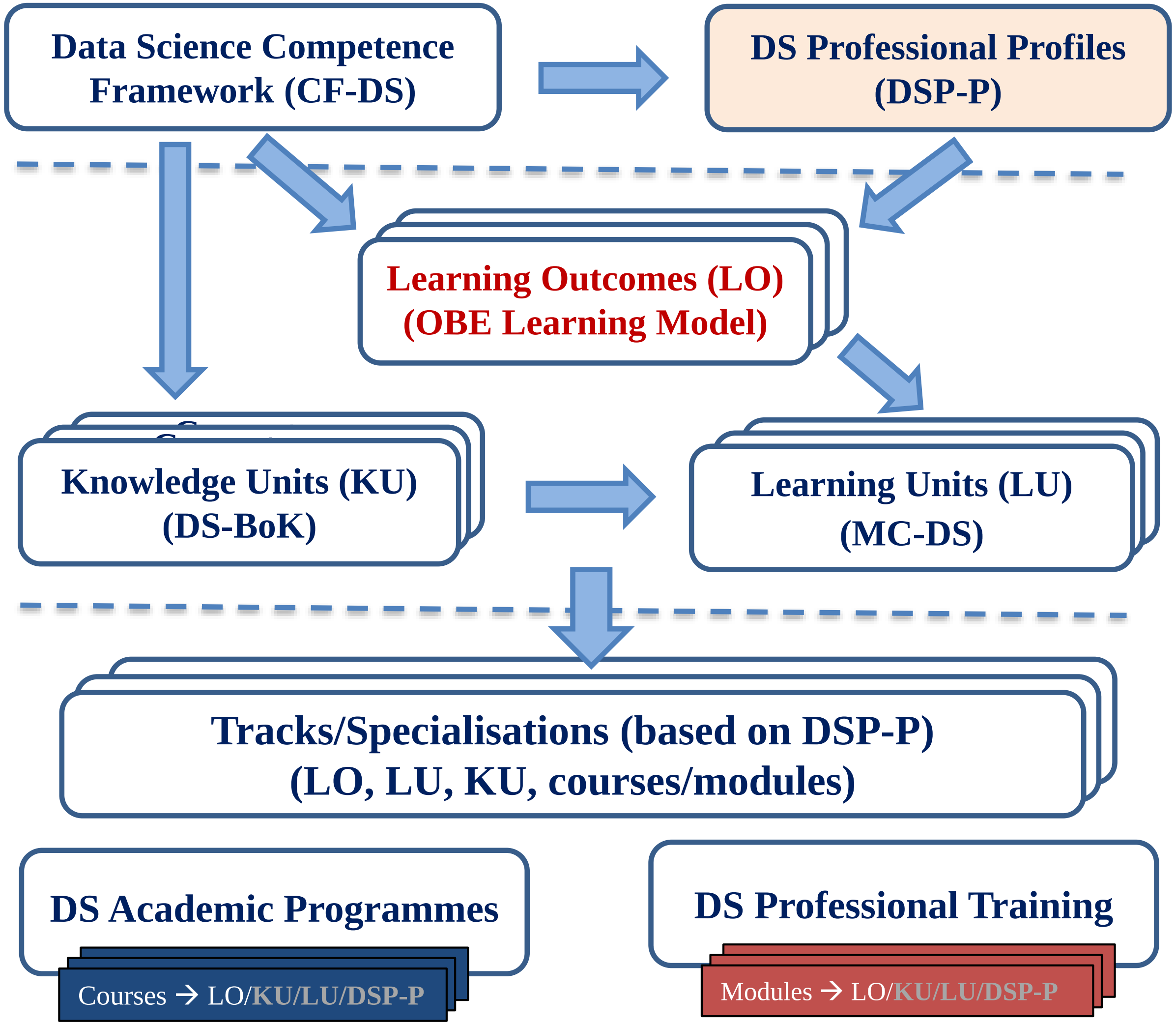
Support workers and data handling clerks: User support workers, data entry clerks, data entry field workers

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

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

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

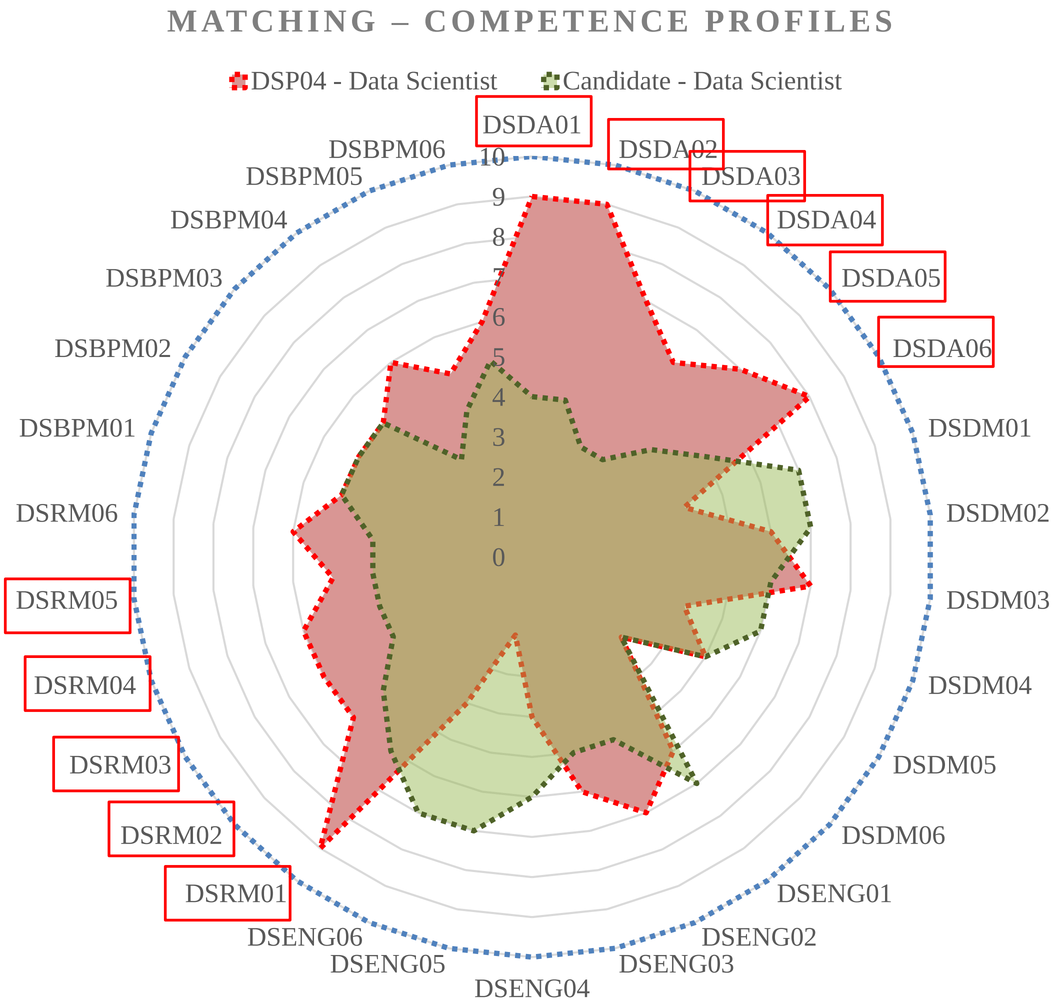
From Competences and DSP Profiles

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

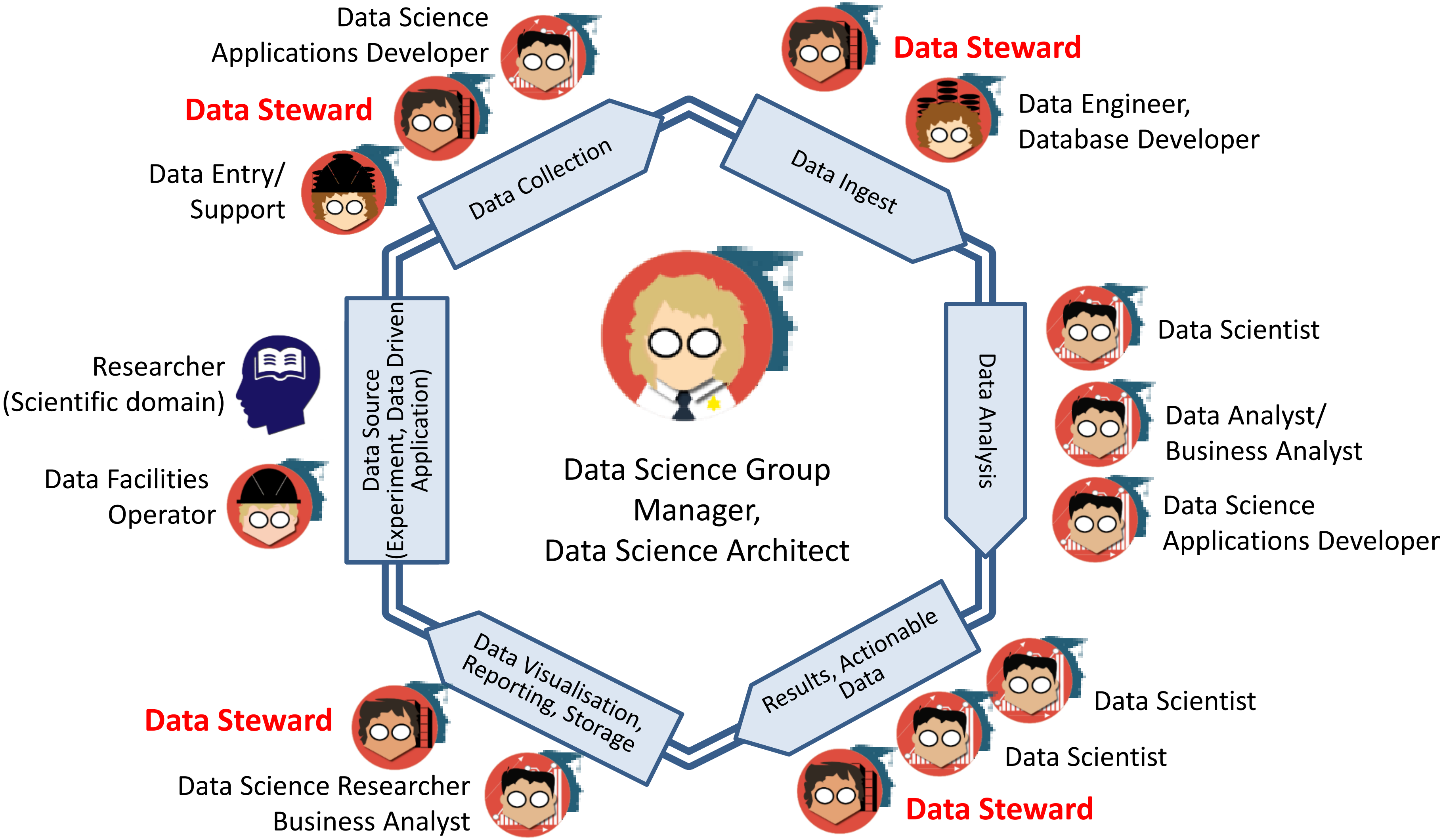


Individual Education/Training Path based on Competence benchmarking

- 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 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 (HLEG EOSC report)
- 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.

EDSF Recognition, Endorsement and Implementation

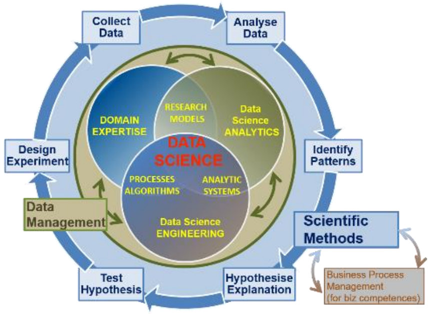
- **DARE (Data Analytics Rising Employment)** project by APEC (Asia Pacific Economic Cooperation)
  - DARE project Advisory Council meeting 4-5 May 2017, Singapore
- **PcW and BHEF Report “Investing in America’s data science and analytics talent”** April 2017
  - Quotes EDSF and Amsterdam School of Data 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

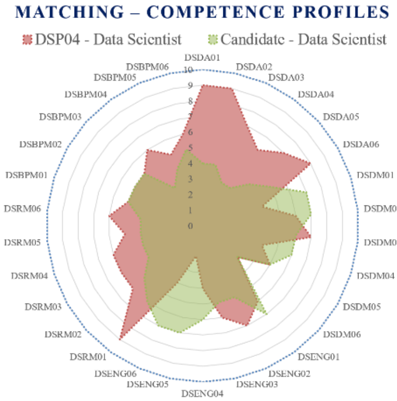
Summary: Services and References

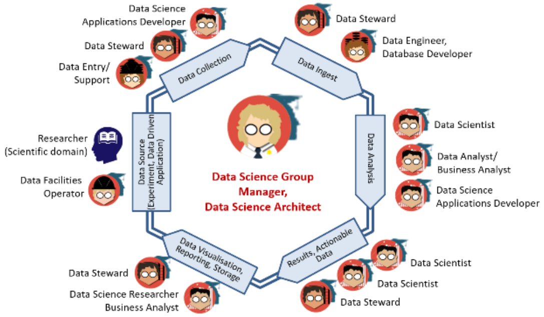
- EDISON Website http://edison-project.eu/
- EDISON Data Science Framework (EDSF) http://edison-project.eu/edison/edison-data-science-framework-edsf
- Directory of University programs http://edison-project.eu/university-programs-list
- Community Portal http://datasciencepro.eu/

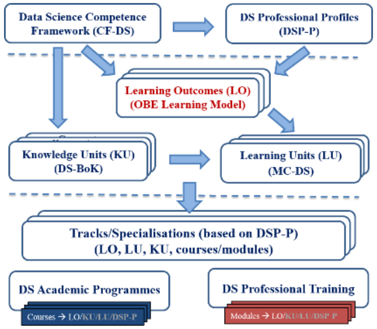
- Survey Data Science Competences: Invitation to participate https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession

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




MATCHING – COMPETENCE PROFILES


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


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

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

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***

21st Century Skills (DARE & BHEF & EDISON)

1. **Critical Thinking:** Demonstrating the ability to apply critical thinking skills to solve problems and make effective decisions
2. **Communication:** Understanding and communicating ideas
3. **Collaboration:** Working with other, appreciation of multicultural difference
4. **Creativity and Attitude:** Deliver high quality work and focus on final result, intitiative, intellectual risk
5. **Planning & Organizing:** Planning and prioritizing work to manage time effectively and accomplish assigned tasks
6. **Business Fundamentals:** Having fundamental knowledge of the organization and the industry
7. **Customer Focus:** Actively look for ways to identify market demands and meet customer or client needs
8. **Working with Tools & Technology:** Selecting, using, and maintaining tools and technology to facilitate work activity
9. **Dynamic (self-) re-skilling:** Continuously monitor individual knowledge and skills as shared responsibility between employer and employee, ability to adopt to changes
10. **Professional networking:** Involvement and contribution to professional network activities
11. **Ethics:** Adhere to high ethical and professional norms, responsible use of power data driven technologies, avoid and disregard un-ethical use of technologies and biased data collection and presentation

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 years
  - EDSF will be maintained by UvA
  - CP by Engineering (Italy)

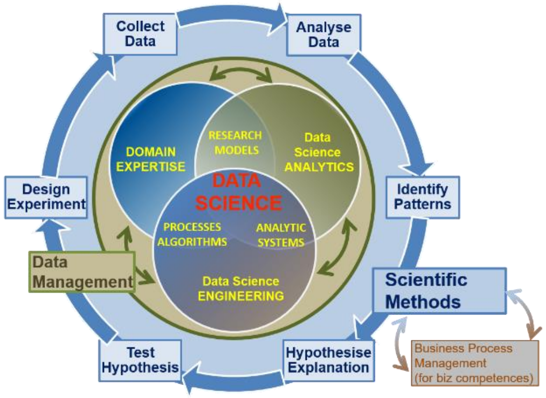
Further developments and Next steps (3)

- 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

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

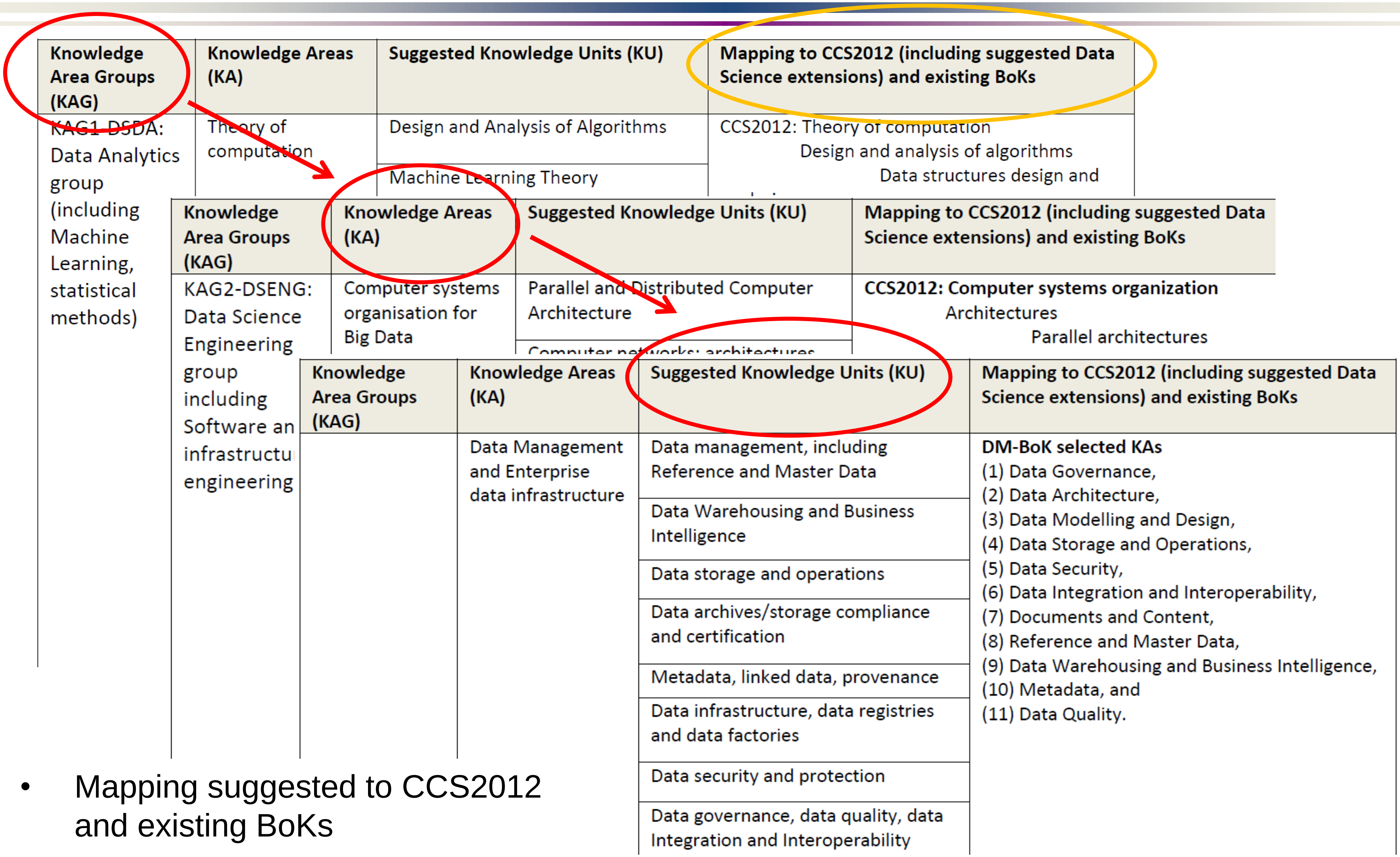
- Data Science domain knowledge to be defined by related expert groups

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)



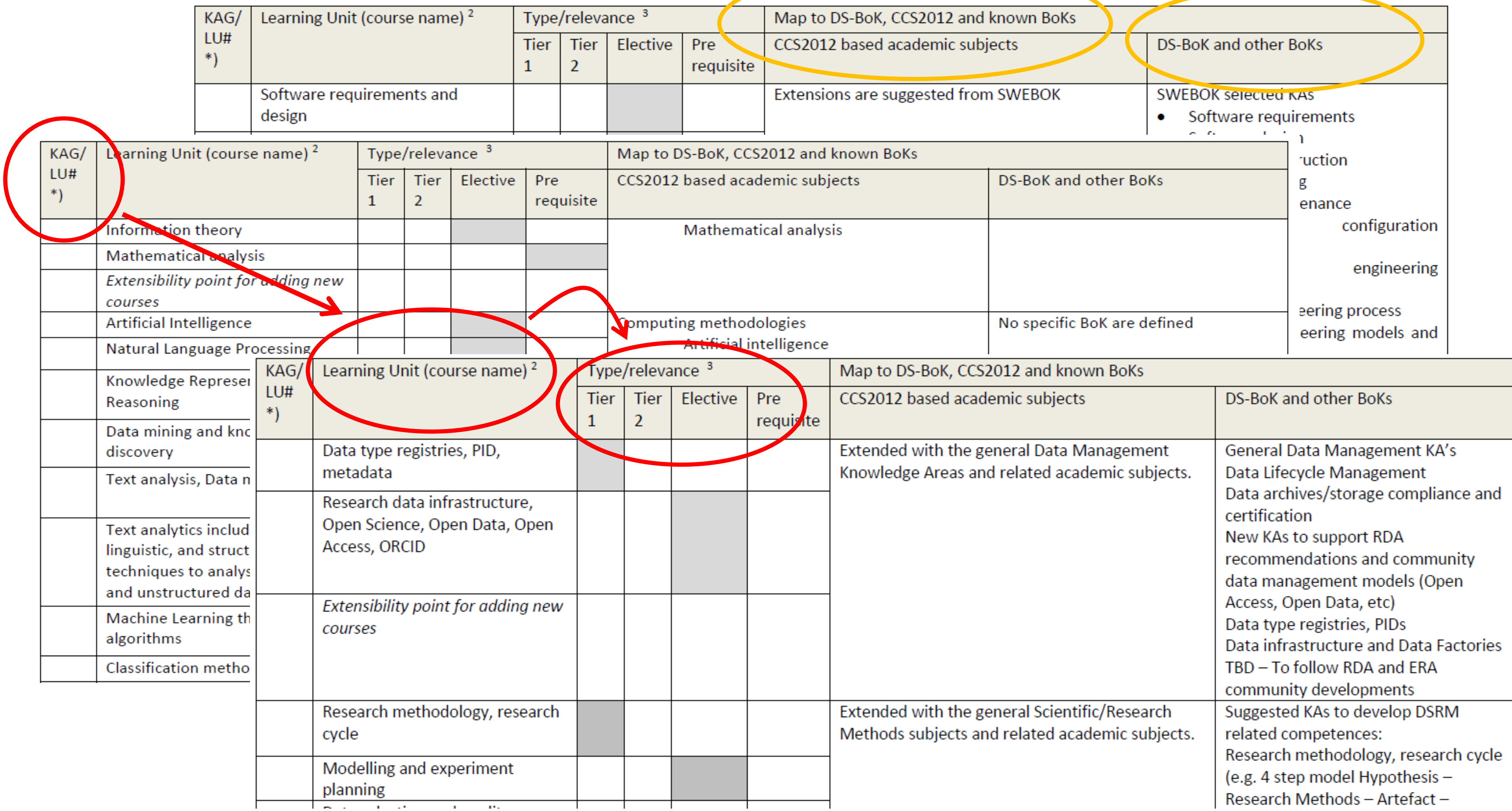
Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Suggested Knowledge Units (KU)	Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs
KAG1-DSDA: Data Analytics group (including Machine Learning, statistical methods)	Theory of computation	Design and Analysis of Algorithms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

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)

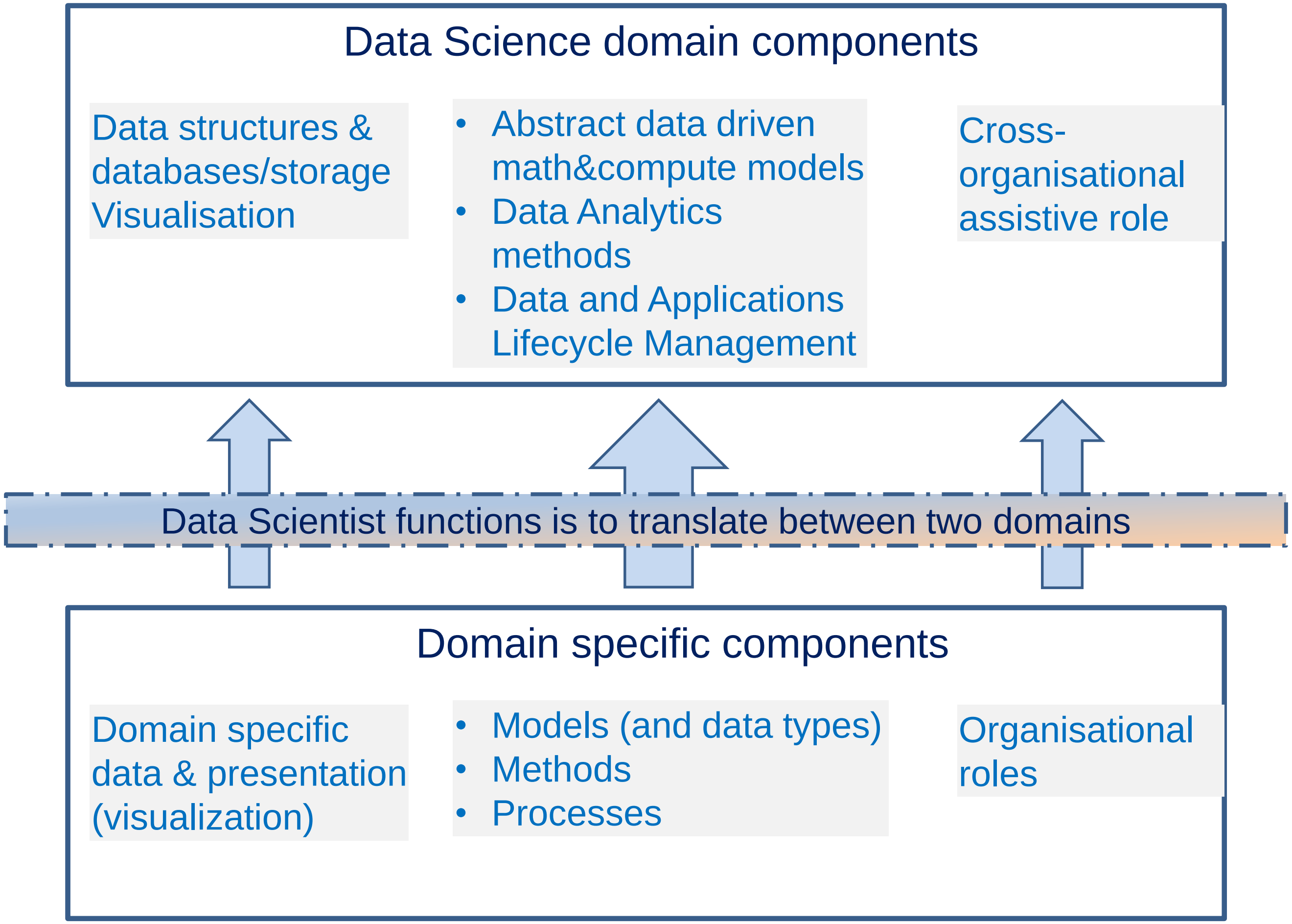
Estimated: Group of 10-12 data specialists for research institution of 200-300 research staff.

Growing role and demand for Data Stewards and data stewardship

Data 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**