Defining a new profession of the Data Scientist to address critical skills gap in European research and industry

Yuri Demchenko, EDISON
University of Amsterdam

Cloud and DevOps World 2016

21-22 June 2016, Olympia, London, UK

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 and motivation
  - Demand for Data Science and data related professions
  - European initiatives related to Digital Single Market (DSM) and demand to data related competences and skills
- EDISON Data Science Framework
  - From Data Science Competences to Body of Knowledge and Model Curriculum
- Data Science Competence Framework: Essential competences and skills
- Taxonomy: Data Science professions family
- Data Science Body of Knowledge (DS-BoK)
  - Knowledge areas and academic disciplines
- Further steps

UvA Universiteit van Amsterdam

University of Southampton

inmark europa

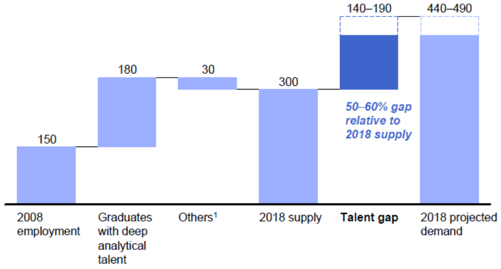
University of Stavanger

egi

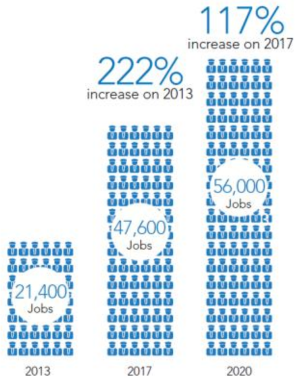
ENGINEERING

FTK Forschungsinstitut für Telekommunikation und Kooperation

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 (2016) identified need for data experts and data stewards
  - Recommendation: Allocate 5% from grant funding for Data management and preservation
  - Estimation: More than 80,000 data stewards (1 per every 20 scientists)
  - Core data experts need to be trained and their career perspective improved

Recent European Commission Initiatives

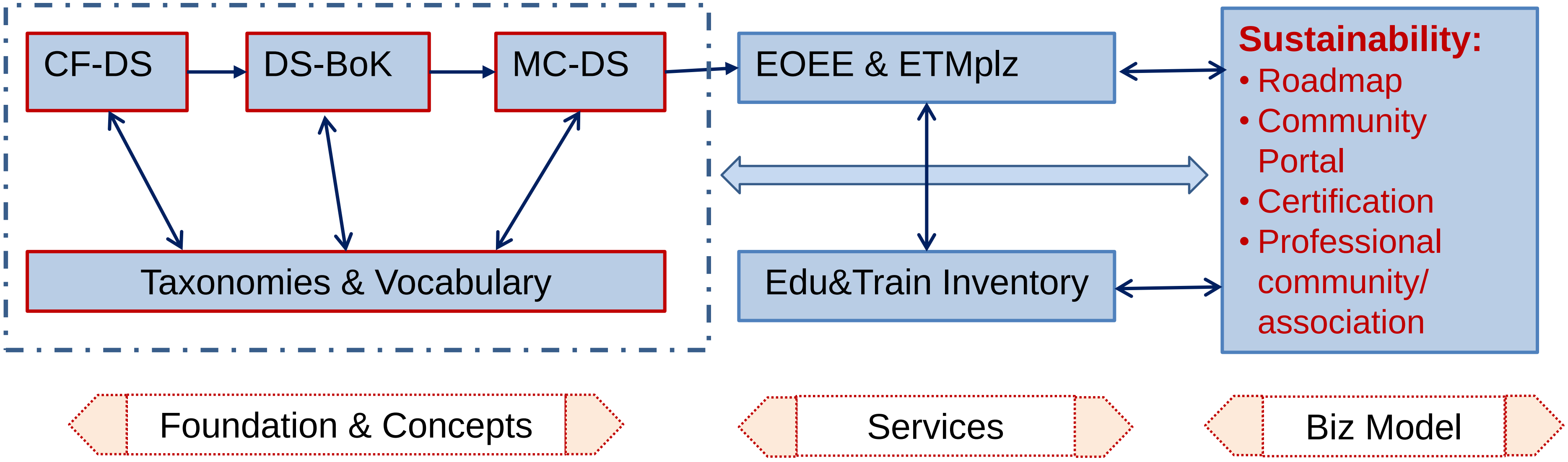
Digitising European Industry: Reaping the full benefits of a **Digital Single Market**. COM(2016) 180 final, Brussels, 19.4.2016

- The need for new multidisciplinary and digital skills is exploding, including such as (Data Scientist) combining data analytics and business or engineering skills.
  - Expected rapidly growing demand will lead to more than 800 000 unfilled vacancies by 2020.
- The forthcoming New Skills Agenda for Europe (exp May 2016) to address the need for digital and complementary skills, ensure young talents flow into data driven research and industry

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
- Raise awareness and change incentive structures for academics, industry and public services to share their data, and improve data management training, literacy and data stewardship skills
- Address growing demand and shortage of data-related skills and lack of recognition of their value (in all sectors).

EDISON Framework: Building the Data Science Profession



- EDISON Framework components
  - CF-DS – Data Science Competence Framework
  - DS-BoK – Data Science Body of Knowledge
  - MC-DS – Data Science Model Curriculum
  - Data Science Taxonomies and Scientific Disciplines Classification
    - Linked to e-CFv3.0, ACM CCS (2012) and ESCO
  - EOEE - EDISON Online Education Environment

Background Frameworks and Standards

- 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
  - Intended inclusion of the Data Science occupations family – end of 2016
- ACM Classification of Computer Science – CCS (2012)
  - ACM Computer Science Body of Knowledge (CS-BoK) and ACM and IEEE Computer Science Curricula 2013 (CS2013)

EDISON Data Science Competence Framework

Data Science Competence Framework is a foundation for the Data Science Profession definition

- How it was made
- 5 main Data Science competences groups
- Skills, tools and languages
- Practical use of the CF-DS (examples)
  - Suggested e-CFv3.0 extensions for Data Science
  - Data Science occupations family

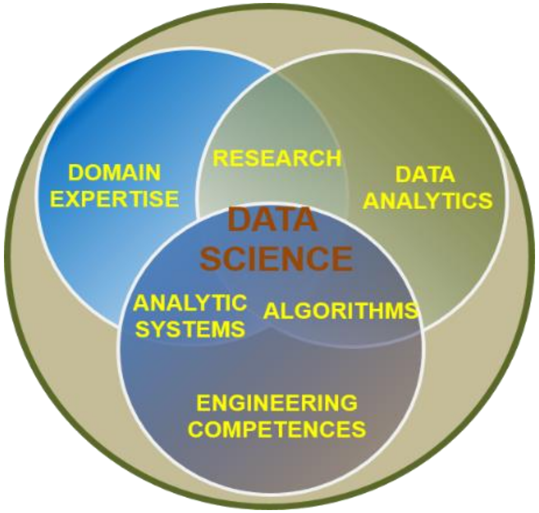
Demanded Data Science Competences and Skills: Jobs market analysis

- Initial Analysis (period Aug – Sept 2015)
  - 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
  - Using manually data analytics methods: classification, clustering, expert evaluation
  - Research methods: Data collection - Hypothesis – Artefact - Evaluation
- Observations
  - Many job ads don’t use Data Scientist as a definite profession:
    - Data Science competences/skills are specified as part of traditional ICT professions/positions
  - Many academic openings without specified skills profile
  - Explicit Data Scientist jobs specify wide variety of expected functions/responsibilities and required skills and knowledge

Data Scientist definition by NIST

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**.*

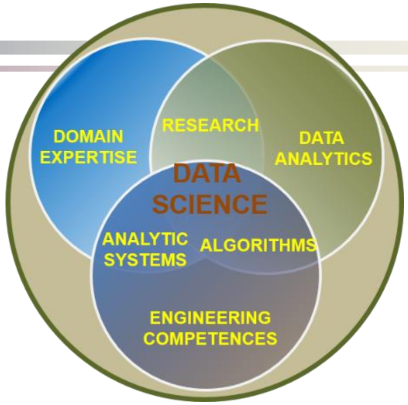


[ref] Legacy: NIST BDWG definition of Data Science

- ***Data science** is the empirical synthesis of actionable knowledge and technologies required to handle data from raw data through the complete data lifecycle process.*

Identified Data Science Competence Groups

- Commonly accepted Data Science competences/skills groups include
  - Data Analytics or Business Analytics or Machine Learning
  - Engineering or Programming
  - Subject/Scientific Domain Knowledge



[ref] Legacy: NIST BDWG definition of Data Science

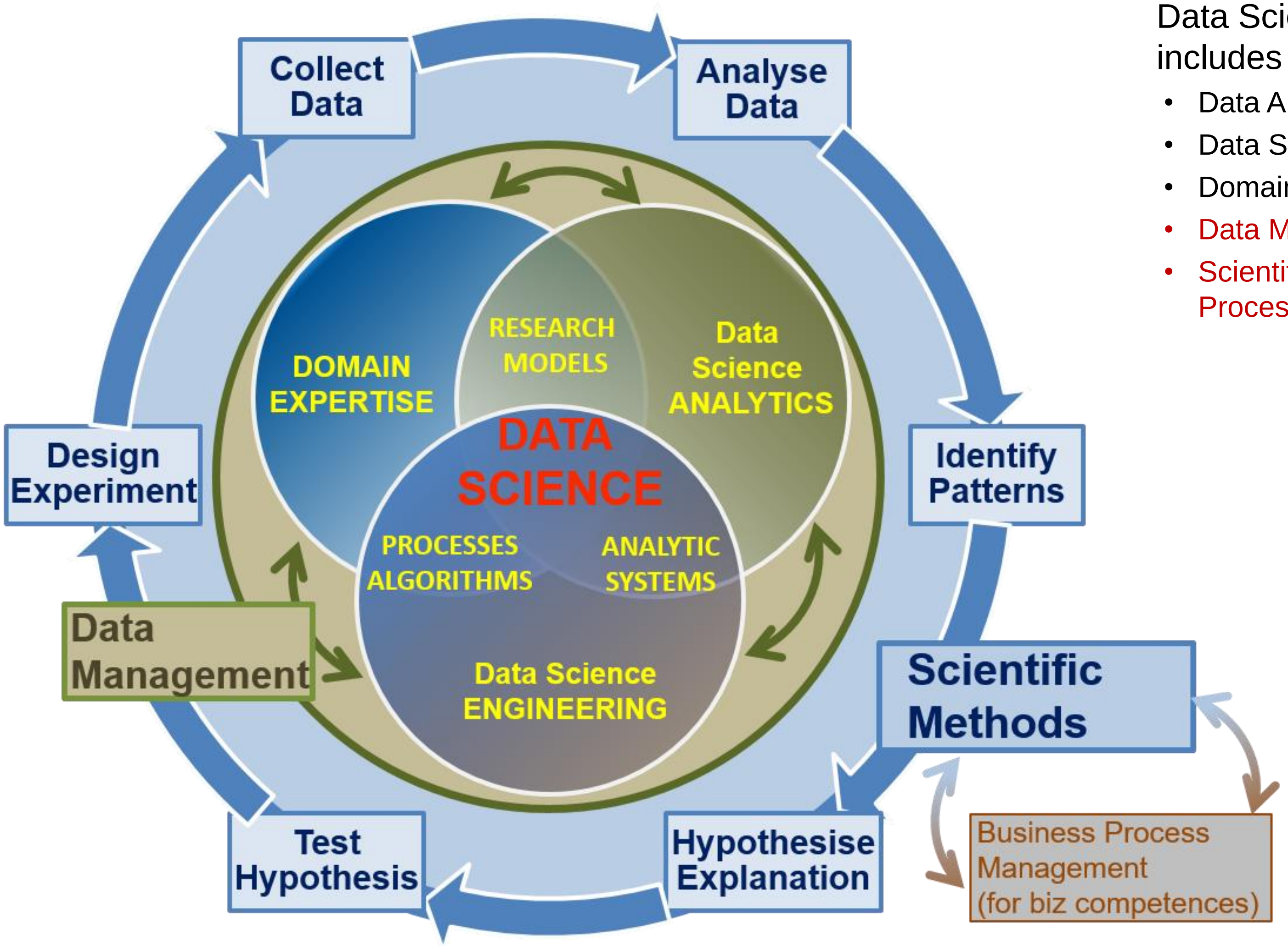
- EDISON identified 2 additional competence groups demanded by organisations
  - **Data Management, Curation, Preservation**
  - **Scientific or Research Methods and/vs Business Processes/Operations**

- Other skills commonly recognized aka “soft skills” or “social intelligence”
  - Inter-personal skills or team work, cooperativeness

- All groups need to be represented in Data Science curriculum and training programmes
  - Challenging task for Data Science education and training

- Another aspect of integrating Data Scientist into organisation structure
  - General Data Science (or Big Data) **literacy** for all involved roles and management
  - Common agreed and understandable way of communication and **information/data presentation**
  - ***Role of Data Scientist: Provide such literacy advice and guiding to organisation***

Data Science Competence Groups - Research



Data Science Competence includes 5 areas/groups

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

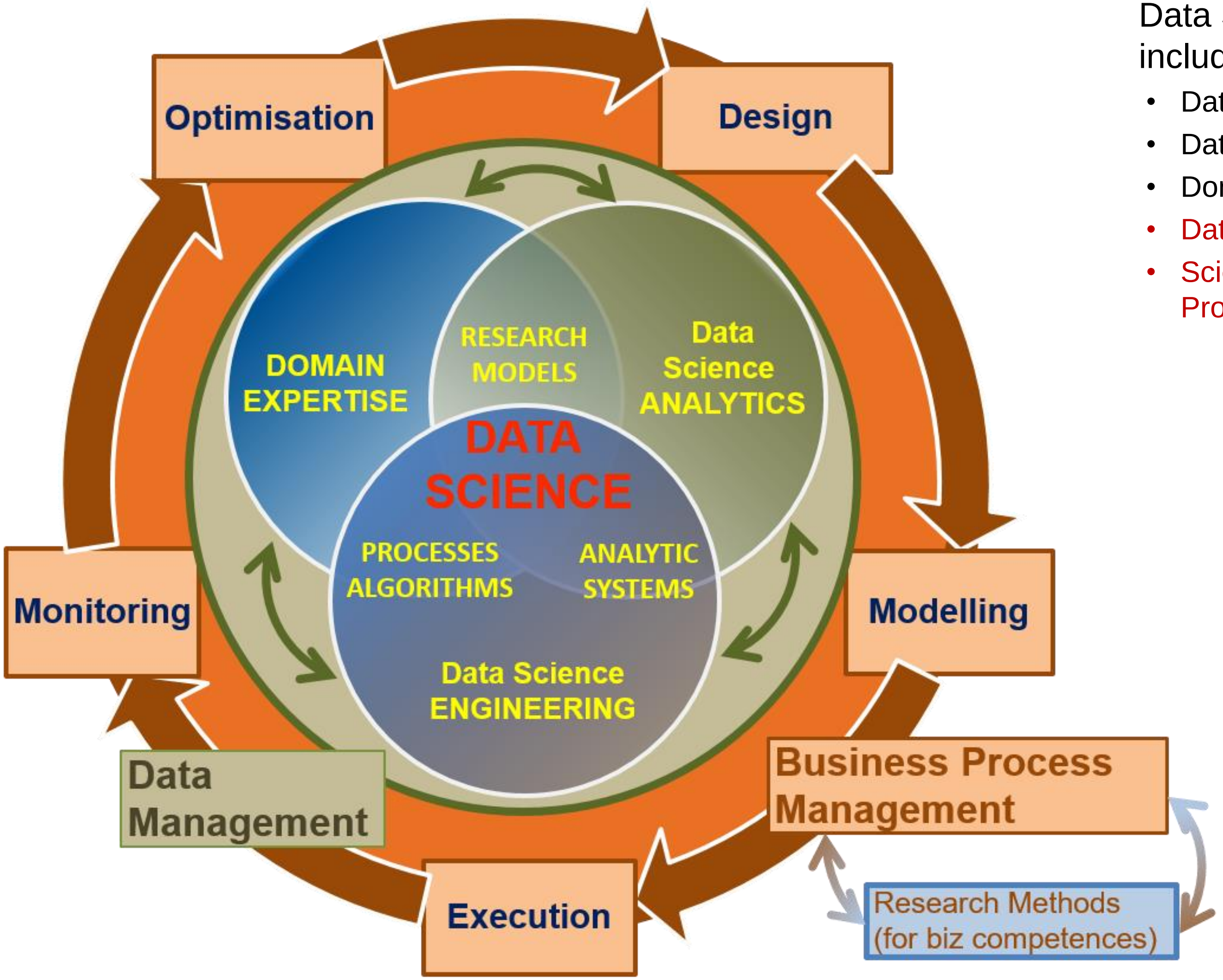
Scientific Methods

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

Business Operations

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

Data Science Competences Groups – Business



Data Science Competence includes 5 areas/groups

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

Scientific Methods

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

Business Process Operations/Stages

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

Identified Data Science Competence Groups (Updated)

	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

Data Science Skill Groups related to Competences

	Data Analytics and Machine Learning	Data Management/ Curation	Data Science Engineering (hardware and software)	Scientific/ Research Methods	Personal/Inter-personal communication, team work	Application/subject domain (research or business), examples
1	Artificial intelligence, machine learning	Manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources	Design efficient algorithms for accessing and analyzing large amounts of data	Analytical, independent, critical, curious and focused on results	Communication skills	Recommender or Ranking system
2	Machine Learning and Statistical Modelling	Data models and datatypes	Big Data solutions and advanced data mining tools	Confident with large data sets and ability to identify appropriate tools and algorithms	Inter-personal intra-team and external communication	Data Analytics for commercial purposes
3	Machine learning solutions and pattern recognition techniques	Experience of working with large data sets	Multi-core/distributed software, preferably in a Linux environment	Flexible analytic approach to achieve results at varying levels of precision	Network of contacts in Big Data community	Data sources and techniques for business insight and customer focus
4	Supervised and unsupervised learning	(non)relational and (un)-structured data	Databases, database systems, SQL and NoSQL	Interest in data science, exceptional analytical skills		Mechanism Design and/or Latent Dirichlet Allocation
5	Data mining	Cloud based data storage and data management	Statistical analysis languages and tooling			Game Theory
6	Markov Models, Conditional Random Fields	Data management planning	Cloud powered applications design			Copyright and IPR
7	Logistic Regression, Support Vector Machines	Metadata annotation and management				
8	Predictive analysis and statistics (including Kaggle platform)	Data citation, metadata, PID (*)				
9	(Artificial) Neural Networks					

Mathematics foundation:
Knowledge of mathematics, calculus, probability theory and statistics

Identified Big Data Tools and Programming Languages

	Big Data Analytics platforms	Math& Stats tools	Databases	Data/ applications visualization	Data Management and Curation platform
1	Big Data Analytics platforms	Advanced analytics tools (R, SPSS, Matlab, etc)	SQL and relational databases	Data visualization Libraries (D3.js, FusionCharts, Chart.js, other)	Data modelling and related technologies (ETL, OLAP, OLTP, etc)
2	Big Data tools (Hadoop, Spark, etc)	Data Mining tools: RapidMiner, others	NoSQL Databases	Visualisation software (D3, Processing, Tableau, Gephi, etc)	Data warehouses platform and related tools
3	Distributed computing tools a plus (Spark, MapReduce, Hadoop, Hive, etc.)	Mathlab	NoSQL, Mongo, Redis	Online visualization tools (Datawrapper, Google Charts, Flare, etc)	Data curation platform, metadata management (ETL, Curator's Workbench, DataUp, MIXED, etc)
4	Real time and streaming analytics systems (like Flume, Kafka, Storm)	Python	NoSQL, Teradata		Backup and storage management (iRODS, XArch, Nesstar, others
5	Hadoop Ecosystem/platform	R, Tableau R	Excel		
6	Spotfire	SAS			
7	Azure Data Analytics platforms (HDInsight, APS and PDW, etc)	Scripting language, e.g. Octave			
8	Amazon Data Analytics platform (Kinesis, EMR, etc)	Statistical tools and data mining techniques			
9	Other cloud based Data Analytics platforms, e.g. HortonWorks, Vertica LexisNexis HPCC System	Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)			

Highlighted:
Cloud based and online data analytics and data management platforms

Suggested 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 (including CV matching)
  - For customizable training and career development
- 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

Suggested e-CF Competences for Data Science:
Next eCF Workshop meeting – Presented 14 April 2016

A. PLAN and Design (9 basic competences)
- A.10* Organisational workflow/processes model definition/formalisation
- A.11* Data models and data structures

B. BUILD: Develop and Deploy/Implement (6 basic competences)
- B.7* Apply data analytics methods (to organizational processes/data)
- B.8* Data analytics application development
- B.9* Data management applications and tools
- B.10* Data Science infrastructure deployment

C. RUN: Operate (4 basic competences)
- C.5* User/Usage data/statistics analysis
- C.6* Service delivery/quality data monitoring

D. ENABLE: Use/Utilise (12 basic competences)
- D10. Information and Knowledge Management (powered by DS)
- D.13* Data presentation/visualisation, actionable data extraction
- D.14* Support business processes/roles with data and insight (support to D.5, D.6, D.7, D.12)
- D.15* Data management/preservation/curation with data and insight

E. MANAGE (9 basic competences)
- E.10* Support Management and Business Improvement with data and insight (support to E.5, E.6)
- E.11* Data analytics for (business) Risk Analysis/Management (support to E.3)
- E.12* ICT and Information security monitoring and analysis (support to E.8)

15 Data Science Competences proposed covering different organizational roles and workflow stages
- Data Scientist roles are crossing multiple org roles and workflow stages

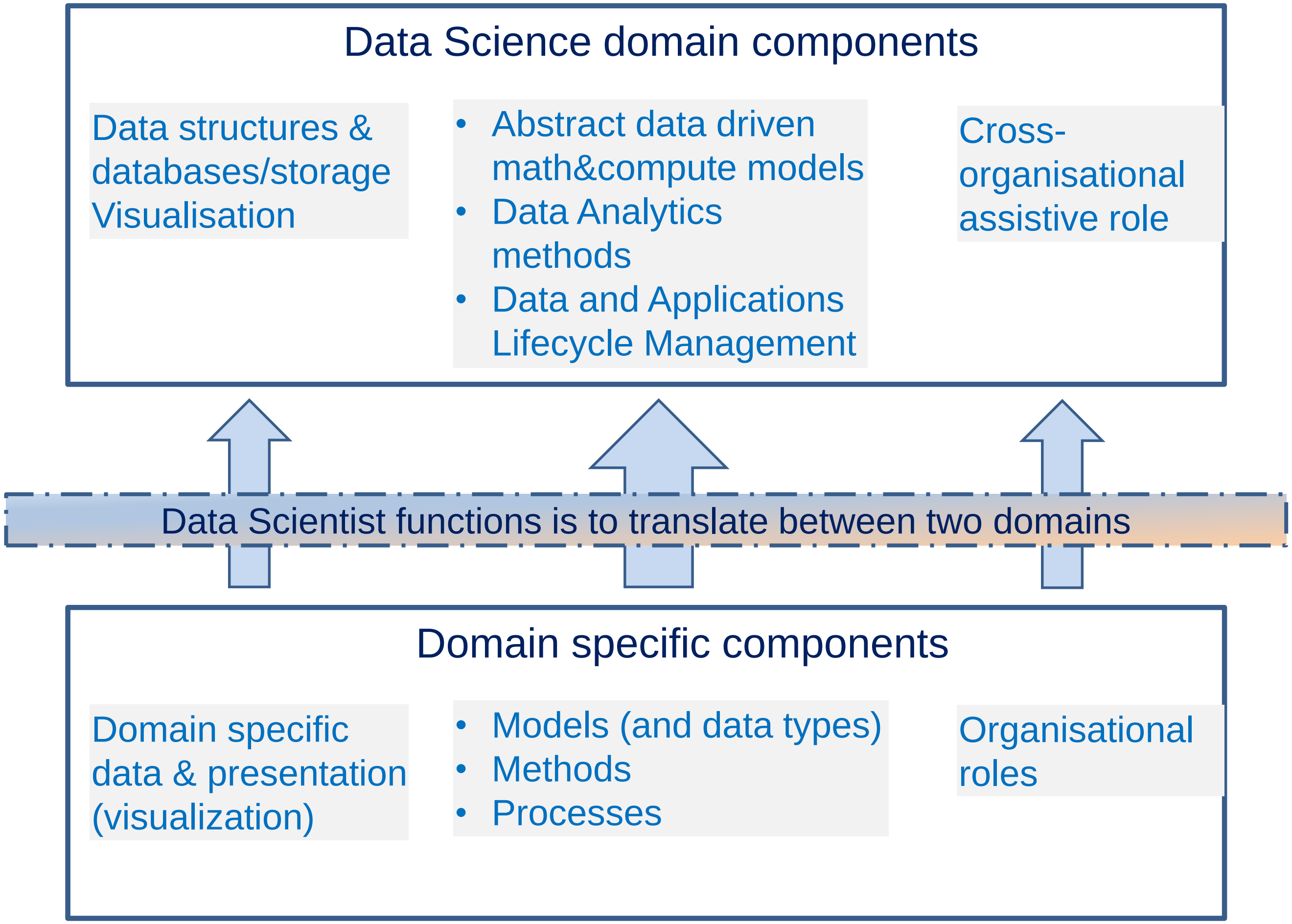
Data Scientist and Subject Domain Specialist

- **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

- **Overall goal: Maintain the Data Value Chain:**
  - Data Integration => Organisation/Process/Business Optimisation => **Innovation**

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)

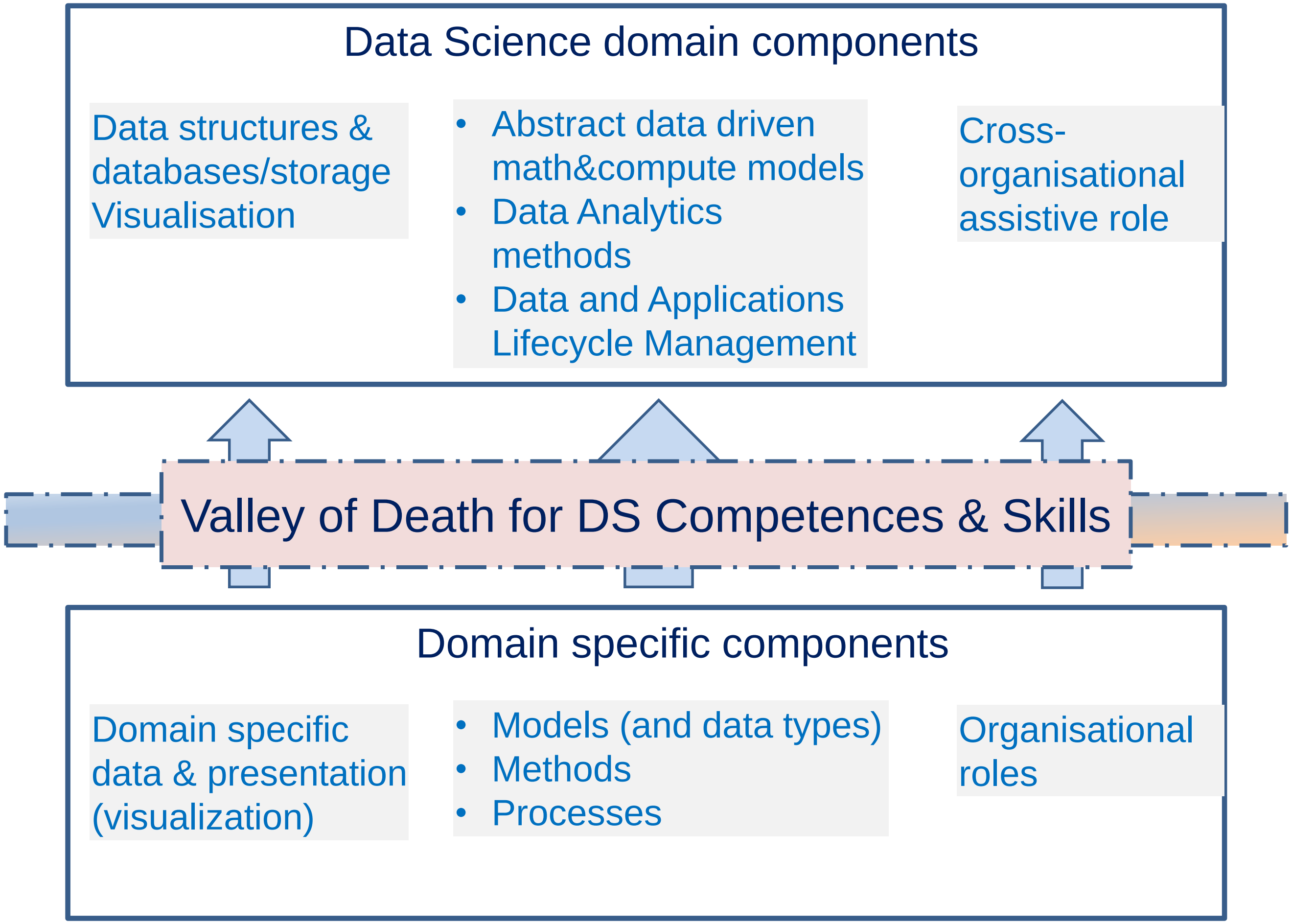
- Models (and data types)
- Methods
- Processes

Organisational roles

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

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

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

Valley of Death for DS Competences & Skills

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**

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

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

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

Data Science Occupations:
Extension for the ESCO (2016) taxonomy (1)

Professionals				
	Science and engineering professionals	**Data Science Professionals**	Data Science professionals not elsewhere classified	DSP04 Data Scientist
				DSP05 Data Science Researcher
				DSP08 (Big) Data Analyst
				DSP07 Data Science (Application) Programmer
				DSP09 Business Analyst
		Database and network professionals	Large scale (cloud) data storage designers and administrators	DSP14 Large scale (cloud) database designer*)
			Database designers and administrators	DSP15 Large scale (cloud) database administrator*)
			Database and network professionals not elsewhere classified	DSP16 Scientific database administrator*)
	Information and communications technology professionals	**Data Science technology professionals**	Data handling professionals not elsewhere classified	DSP12 Digital Librarian
				DSP13 Data Archivist
				DSP10 Data Steward
				DSP11 Data curator

19 DSP# Enumerated Data Science profiles defined by EDISON Framework

Data Science Occupations: Extension for the ESCO taxonomy (2)

Technicians and associate professionals				
	Science and engineering associate professionals	**Data Science Technology Professionals**	Data Infrastructure engineers and technicians	DSP17 Big Data facilities Operators
				DSP18 Large scale (cloud) data storage operators
			Database and network professionals not elsewhere classified	DSP19 Scientific database operator*)
Managers				
	Production and specialised services managers	**Data Science/Big Data Infrastructure Managers**		DSP01/DSP02 Data Science/Big Data Infrastructure Manager
			Research Infrastructure Managers	DSP03 RI Manager
				DSP03 RI Data storage facilities manager
Clerical support workers				
	General and keyboard clerks			
	Data handling support workers (alternative)	**Data and information entry and access**	Digital Archivists and Librarians	DSP12* Digital Librarian
				DSP13* Data Archivist
				DSP10* Data Steward
				DSP11* Data curator

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
- Maintaining Data Value Chain

- Group Manager
- Data Science Architect
- Data Analyst
- Data Science Application programmer
- Data Infrastructure/facilities administrator/operator: storage, cloud, computation
- Data stewards

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
- Maintaining Data Value Chain

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

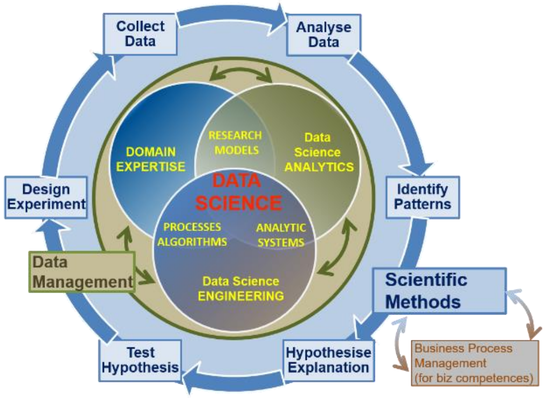
Estimated: Group of 10-12 specialists for research institution of 200-300 staff.

Education and Training

- Foundation and methodological base
  - Data Science Body of Knowledge (DS-BoK)
    - Taxonomy and classification of Data Science related scientific subjects
  - Instructional methodologies and teaching models
- Platforms and environment
  - Virtual labs, datasets, developments platforms
  - Access control and accounts/identity management
  - Online education environment and courses management
- Services
  - Individual benchmarking and profiling tools (competence assessment)
  - Knowledge evaluation tools
  - Certifications and training
  - Education and training marketplace: Courses catalog and repository

Data Science Body of Knowledge (DS-BoK)

DS-BoK Knowledge Area Groups (KAG)



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

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

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

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:

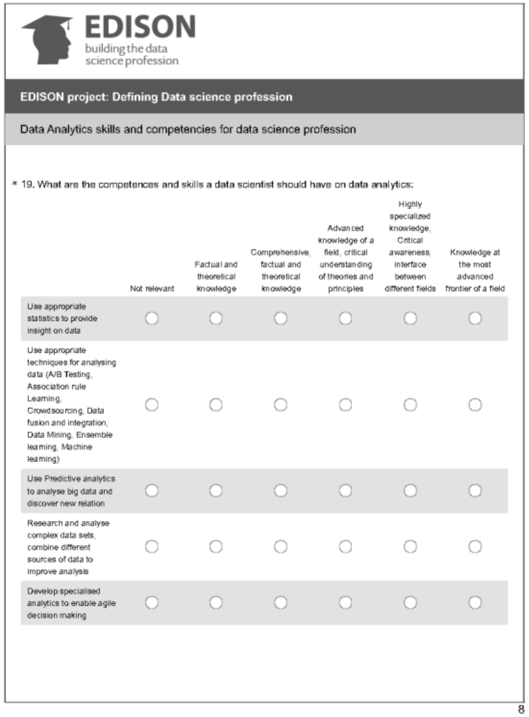
a) Data Management Plan design
b) Metadata and tools
c) Selection of licenses for open data and contents (e.g. Creative Common and Open Database)

Further Steps

- Define a taxonomy and classification for DS competences and skills as a basis for more formal CF-DS definition
- Run surveys for target communities https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession
  - Plan a number of key interviews, primarily experts and top executives at universities and companies
- Formally provide suggestions for e-CF3.0 extensions for Data Science to CEN/PC 428
  - Involve national e-CF bodies and adopters where available
- Provide feedback and contribution to ESCO with the definition of the Data Science professions family
- Suggest ACM CCS2012 Classification extensions and officially contact ACM
- Involve academic and industry experts and professional organisations in the definition of DS-BoK following from CF-DS
  - Link to taxonomy of academic and educational and training courses
- Invite and analyse community contribution main EDISON deliverables
  - CF-DS, DS-BoK, DS Professional Profiles documents are on public comments and available from the EDISON website

Discussion

- Questions
- Observations
- Suggestions



EDISON
building the data science profession

EDISON project: Defining Data science profession

Data Analytics skills and competencies for data science profession

* 19. What are the competences and skills a data scientist should have on data analytics:

	Not relevant	Factual and theoretical knowledge	Comprehensive, factual and theoretical knowledge	Advanced knowledge of a field, critical understanding of theories and principles	Highly specialized knowledge, Critical awareness, interface between different fields	Knowledge at the most advanced frontier of a field
Use appropriate statistics to provide insight on data	○	○	○	○	○	○
Use appropriate techniques for analysing data (A/B Testing, Association rule Learning, Crowdsourcing, Data fusion and integration, Data Mining, Ensemble learning, Machine learning)	○	○	○	○	○	○
Use Predictive analytics to analyse big data and discover new relation	○	○	○	○	○	○
Research and analyse complex data sets, combine different sources of data to improve analysis	○	○	○	○	○	○
Develop specialised analytics to enable agile decision making	○	○	○	○	○	○

8

- Survey Data Science Competences [1]: Invitation to participate
  https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession

- Community discussion documents: Request for comments
  - 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 Professional Profiles
    http://edison-project.eu/data-science-professional-profiles

Definitions (according to e-CFv3.0)

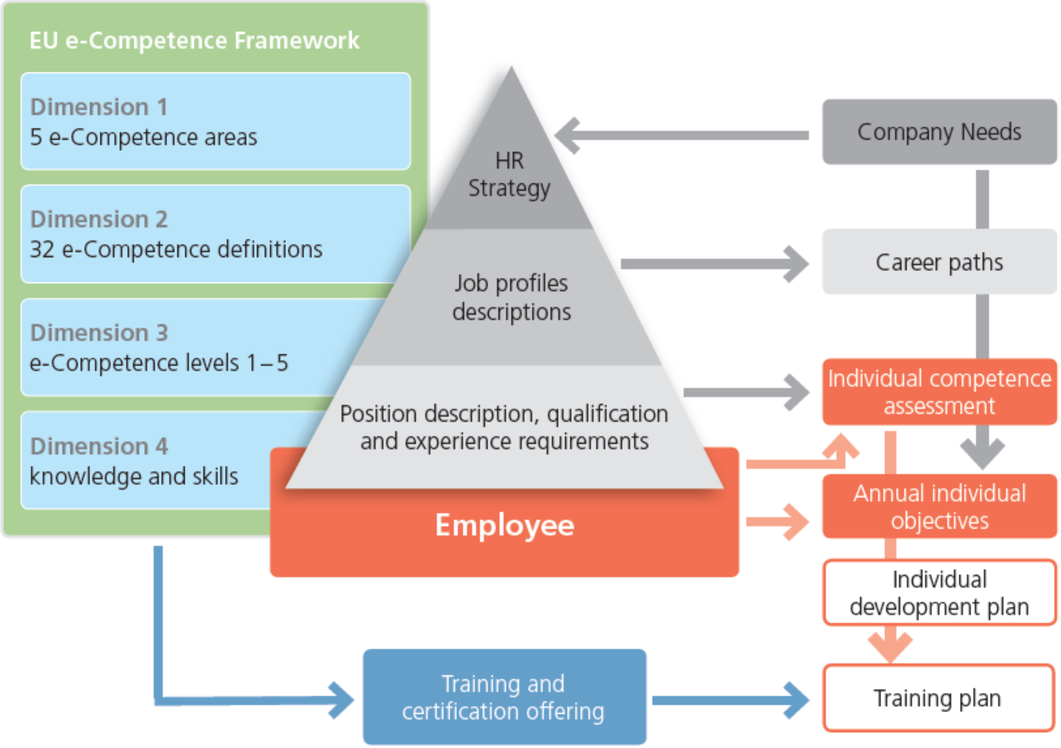
- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results.
  - Competence vs Competency (e-CF vs ACM)
    - Competence is ability acquired by training or education (linked to learning outcome)
    - Competency is similar to skills or experience (acquired feature of a person)
- Competence is not to be confused with process or technology concepts such as, ‘Cloud Computing’ or ‘Big Data’. These descriptions represent evolving technologies and in the context of the e-CF, they may be integrated as elements within knowledge and skill examples.

- **Knowledge** in the context of competence definition is treated as something to know, to be aware of, familiar with, and obtained as a part of education.

- **Skills** is treated as provable ability to do something and relies on the person’s experience.

EDISON Approach: CF-DS and e-CFv3.0

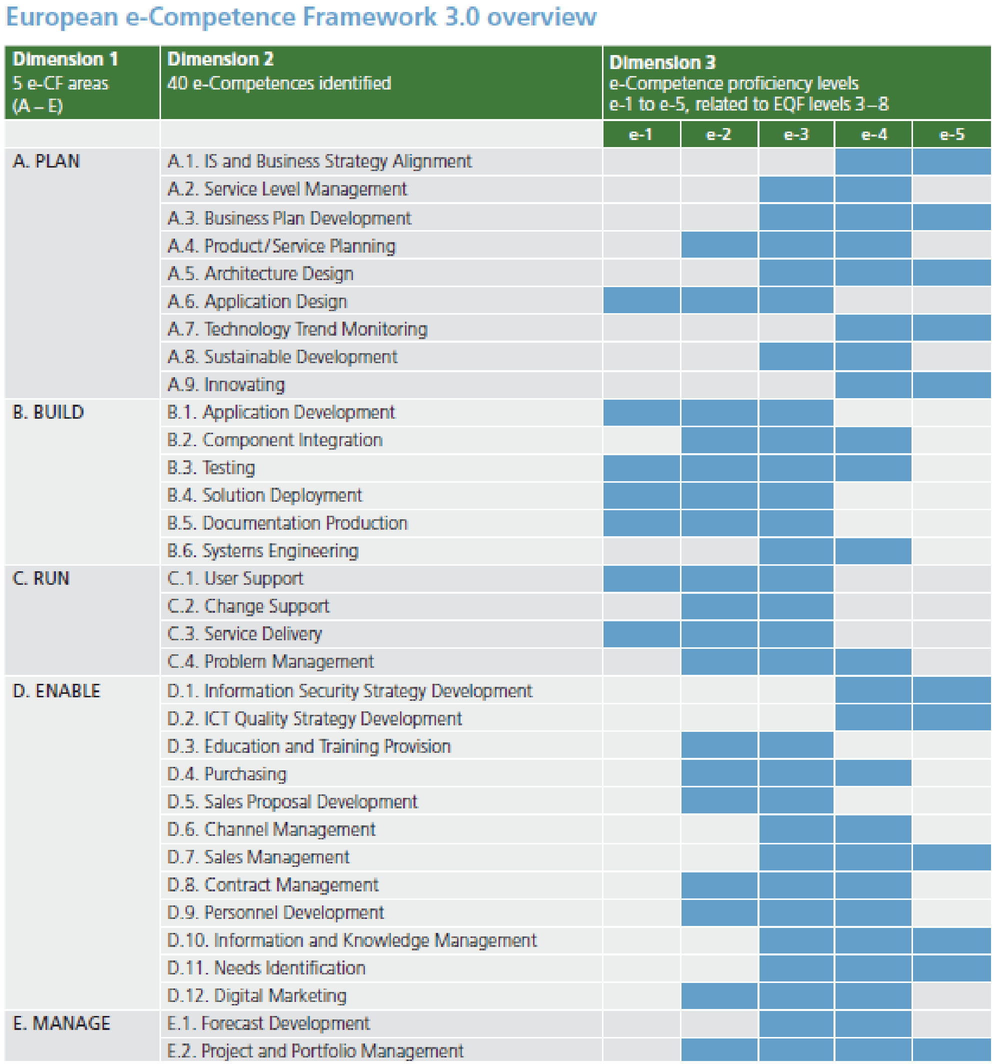
- Competence Framework for Data Science (CF-DS) definition will be built based on European e-Competence framework for IT (e-CFv3.0)
  - Linking *scientific research lifecycle*, organizational roles, competences, skills and knowledge
  - Defining *Data Science Body of Knowledge (DS-BoK)*
  - Mapping CF-DS and DS-BoK to academic disciplines in a DS *Model Curriculum (MC-DS)*


EU e-Competence Framework
Dimension 1
5 e-Competence areas
Dimension 2
32 e-Competence definitions
Dimension 3
e-Competence levels 1 – 5
Dimension 4
knowledge and skills
HR Strategy
Job profiles descriptions
Position description, qualification and experience requirements
Employee
Company Needs
Career paths
Individual competence assessment
Annual individual objectives
Individual development plan
Training plan
Training and certification offering

- Multiple use of e-CFv3.0 within ICT organisations
- Provides basis for individual career path, competence assessment, training and certification

- EDISON CF-DS will be used for defining DS-BoK and MC-DS, linking organizational functions and required knowledge
- Provide basis for individual (self) training and certification

e-CFv3.0 Internal Structure: Refactoring for CF-DS



European e-Competence Framework 3.0 overview

Dimension 1 5 e-CF areas (A – E)	Dimension 2 40 e-Competences identified	Dimension 3 e-Competence proficiency levels e-1 to e-5, related to EQF levels 3–8				
		e-1	e-2	e-3	e-4	e-5
A. PLAN	A.1. IS and Business Strategy Alignment				■	■
	A.2. Service Level Management			■	■	
	A.3. Business Plan Development			■	■	■
	A.4. Product/Service Planning		■	■	■	
	A.5. Architecture Design			■	■	■
	A.6. Application Design	■	■	■		
	A.7. Technology Trend Monitoring				■	■
	A.8. Sustainable Development			■	■	
	A.9. Innovating				■	■
B. BUILD	B.1. Application Development	■	■	■		
	B.2. Component Integration		■	■	■	
	B.3. Testing	■	■	■	■	
	B.4. Solution Deployment	■	■	■		
	B.5. Documentation Production	■	■	■		
	B.6. Systems Engineering			■	■	
C. RUN	C.1. User Support	■	■	■		
	C.2. Change Support		■	■		
	C.3. Service Delivery	■	■	■		
	C.4. Problem Management		■	■	■	
D. ENABLE	D.1. Information Security Strategy Development				■	■
	D.2. ICT Quality Strategy Development				■	■
	D.3. Education and Training Provision		■	■		
	D.4. Purchasing		■	■	■	
	D.5. Sales Proposal Development		■	■		
	D.6. Channel Management			■	■	
	D.7. Sales Management			■	■	■
	D.8. Contract Management		■	■	■	
	D.9. Personnel Development		■	■	■	
	D.10. Information and Knowledge Management			■	■	■
	D.11. Needs Identification			■	■	■
	D.12. Digital Marketing		■	■	■	
E. MANAGE	E.1. Forecast Development			■	■	
	E.2. Project and Portfolio Management		■	■	■	■

- 4 Dimensions
  - Competence Areas
  - Competences
  - Proficiency levels
  - Skills and Knowledge

- 5 Competence Area defined by ICT Business Process stages
  - Plan
  - Build
  - Run
  - Enable
  - Manage

⇒ Consider for Data Science:

- Scientific Research cycle/workflow (or Scientific Data Lifecycle) vs Organisational/project flow

- Each competence has 5 proficiency level
  - Ranging from technical to engineering to management to strategist/expert level

- Knowledge and skills property are defined for/by each competence and proficiency level (not unique)