H2020 INFRASUPP-4 CSA Project EDISON

Defining the Data Science Competence Framework

Overview of Existing Studies and Proposed Approach

Yuri Demchenko, Adam Belloum
University of Amsterdam

EGI Community Forum 10-13 November 2015
Bari, Italy

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

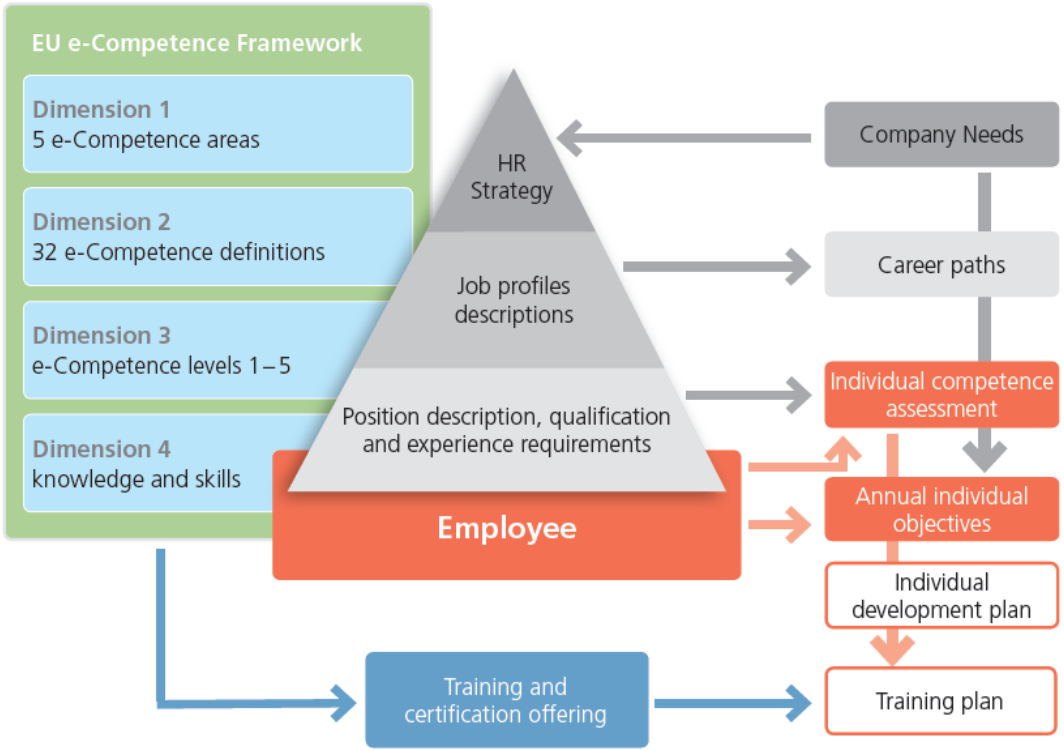
Grant 675419 (INFRASUPP-4-2015: CSA)

Outline

- EDISON approach
  - From Data Science Competences to Body of Knowledge and Model Curriculum
- e-CF3.0 overview and analysis
- Data Science essential skills required
  - Demand side and job market analysis
- Organisational workflow/processes and role of Data Scientist
- Further steps - Survey and questionnaires

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

- 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



- 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 Areas** defined by ICT Business Process stages
  - Plan
  - Build
  - Deploy
  - Run
  - Manage

-> Refactor to Scientific Research (or Scientific Data) Lifecycle

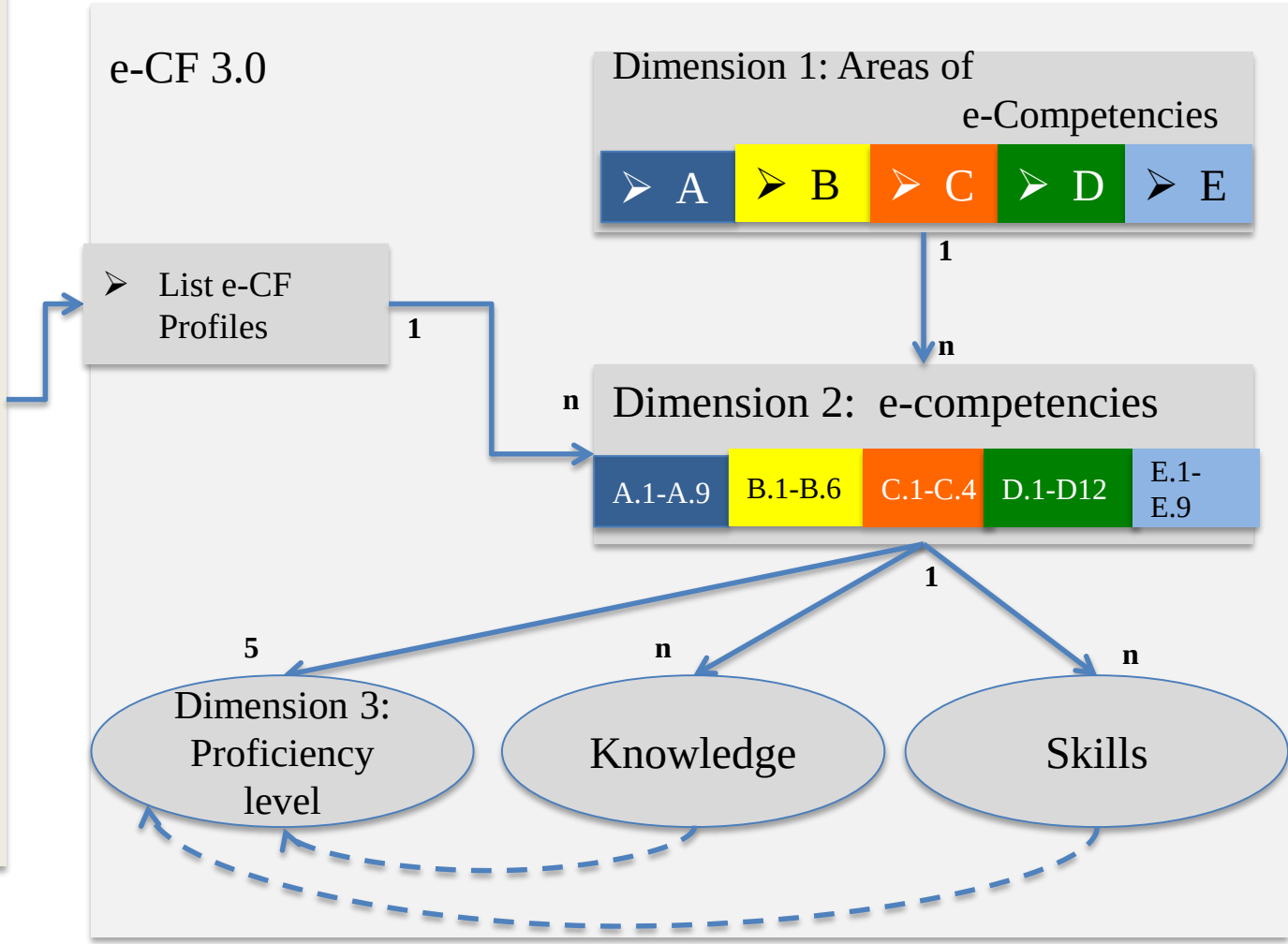
  - See example of RI manager at IG-ETRD wiki and meeting

- Each competence has 5 proficiency levels
  - 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)

EDISON CF-DS profile(s) and e-CF3.0

Edison Profile(s)
For Data Science

1. Define **CF-DS profile** using input from
   1. Demand/Jobs market
   2. Surveys, Interview
   3. Questionnaires
   4. DS programmes
2. Map required background ICT competences from e-CF3.0 and ICT profiles
3. Identify required extensions to e-CF3.0



Definitions (according to eCFv3.0)

- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results.
  - Competence vs Competency
  - Competency is similar to skills or experience
- 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.

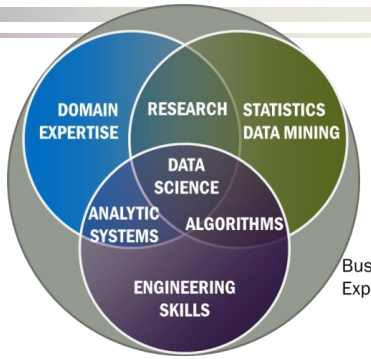
- **Knowledge** in the context of competence definition is treated as something to know, to be aware of, familiar with, and obtained as a part of education.

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

Demanded Data Science Competences and Skills: Jobs market analysis

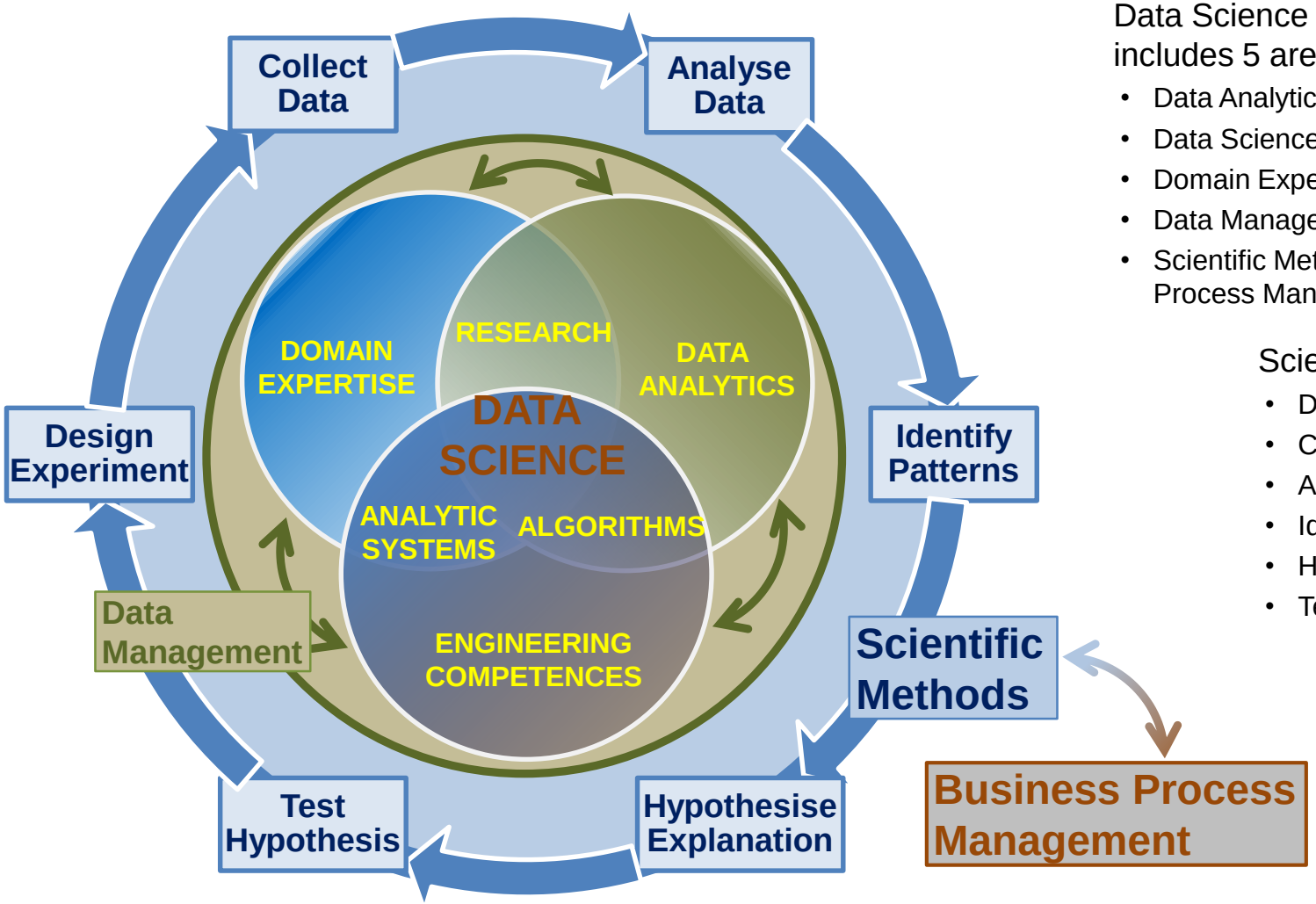
- Source
  - IEEE Data Science Jobs (World but majority US) (collected > 120, selected for analysis > 30)
  - LinkedIn Data Science Jobs (NL) (collected > 140, selected for analysis > 30)
  - Existing studies and reports
- Observations
  - Many job ads don't use Data Scientist as a definite profession:
    - 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

Identified Data Science Competence Groups



- Traditional/known Data Science skills/knowledge profiles include
  - Data Analytics or Business Analytics or Machine Learning
  - Engineering or Programming
  - Subject/Scientific Domain Knowledge
- EDISON identified 2 additional competence groups demanded by organisations
  - Data Management, Curation, Preservation
  - Scientific or Research Methods and/vs Business Operations/Processes
- 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
  - 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 way of communication and information/data presentation
  - *Role of Data Scientist: Be ready to provide such literacy advice and guiding to organisation*

Data Science Competences Areas



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

Identified Data Science Competence Groups

	Data Analytics (DA)	Data Management/ Curation (DM)	DS Engineering (DSE)	Ssearch Methods (DSRM) cientific/Re	DS Domain Knowledge (including Business Apps)
1	Use appropriate statistical techniques on available data to deliver insights	**Develop and implement data strategy**	Use engineering principles to research, design, or develop structures, instruments, machines, experiments, processes, systems, theories, or technologies	Create new understandings and capabilities by using the scientific method's hypothesis, test, and evaluation techniques; critical review; or similar engineering research and development methods	Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
2	Use predictive analytics to analyse big data and discover new relations	**Develop data models including metadata**	Develops specialized data analysis tools to support executive decision making	Direct systematic study toward a fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts, and discovers new approaches to achieve goals	Use data to improve existing services or develop new services
3	Research and analyze complex data sets, combine different sources and types of data to improve analysis.	**Integrate different data source and provide for further analysis**	Design, build, operate relational non-relational databases	Undertakes creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	Participate strategically and tactically in financial decisions that impact management and organizations
4	Develop specialized analytics to enable agile decision making	**Develop and maintain a historical data repository of analysis**	Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	Apply ingenuity to complex problems, develop innovative ideas	Recommends business related strategic objectives and alternatives and implements them
5		**Collect and manage different source of data**	Develop solutions for secure and reliable data access	Ability to translate strategies into action plans and follow through to completion.	Provides scientific, technical, and analytic support services to other organisational roles
6		**Visualise complex and variable data.**	Develop algorithms to analyse multiple source of data	Influences the development of organizational objectives	Analyse multiple data sources for marketing purposes
7			Prototype new data analytics applications		Analyse customer data to identify/optimise customer relations actions

Suggested e-CF extensions for DS

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

B. BUILD: Develop and Deploy/Implement
- 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
- C.5* User/Usage data/statistics analysis
- C.6* Service delivery/quality data monitoring

D. ENABLE: Use/Utilise
- 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

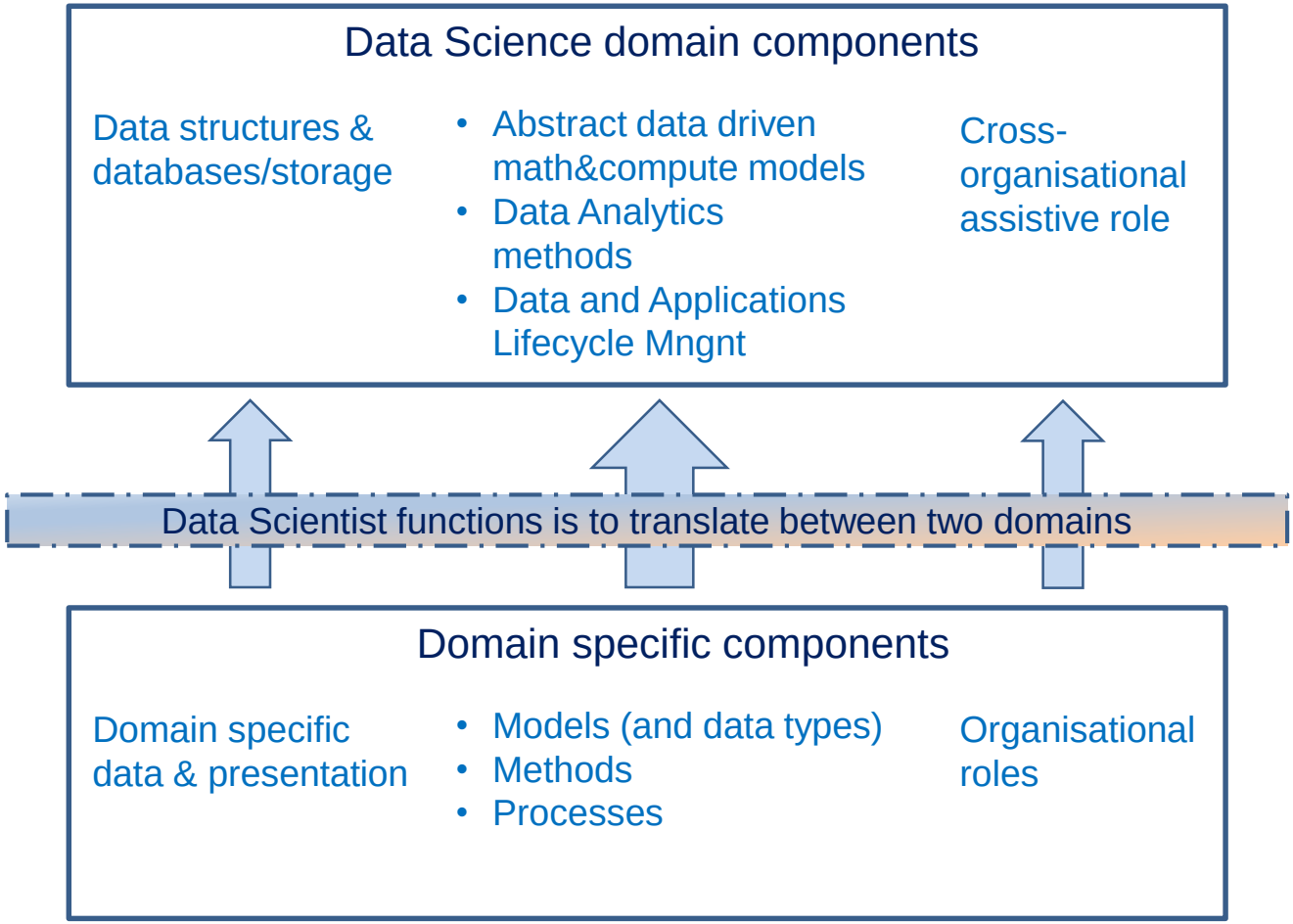
E. MANAGE
- E.10* Support Management and Business Improvement with data and insight (support to E.5, E.6)
- E.11* Data analytics for (business) Risk Analysis/Management (support to E.3)
- E.12* ICT and Information security monitoring and analysis (support to E.8)

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
  - Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data

Data Science and Subject Domains



Possible Data Scientist profiles/roles

- Data Analytics
  - Data Mining
  - Machine Learning
- Data Management
  - Digital Librarian, Data Archivist, Data Curator
- Data Science Engineering
  - Data Analytics applications development
  - Scientific programmer
  - Data Science/Big Data Infrastructure engineer/developer/operator
- Data Science Researcher
  - Data Science creative
  - Data Science consultant/Analyst
- Business Analyst
- Data Scientist in subject/research domain

- Research e-Infrastructure brings its own specifics to required competences and skills definition

Further Steps

- Define a taxonomy and classification for DS competences and skills as a basis for more formal CF-DS definition
  - Cooperate with other H2020 projects
- Create a Questionnaire using CF-DS vocabulary
  - Run surveys for target communities
    - First of all, for EGI community
    - Create open community forum to collect contribution
  - Plan a number of key interviews, primarily experts and top executives at universities and companies