Defining the Data Science Competence Framework (CF-DS)

Overview of Existing Studies and Proposed Approach

EDISON
building the data science profession

Yuri Demchenko
University of Amsterdam

CEN e-CF Workshop @AFNOR
9 December 2015, Paris

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

UvA UNIVERSITEIT VAN AMSTERDAM

UNIVERSITY OF Southampton

inmark europa

University of Stavanger

egi

ENGINEERING

FTK Forschungsinstitut für Telekommunikation und Kooperation

EDISON Objectives

Objective 1: Data Science Curricula Foundation

Promote the creation of *Data Scientists curricula* by an increasing number of universities and professional training organisations.

- The ***Data Science Competence Framework (CF-DS)*** including Taxonomy of competences and skills, compliant with e-CF3.0.
- The ***Body of Knowledge (BoK) for the Data Science (DS-BoK)*** that will map required competencies/skills and existing academic, research and technology disciplines
- A ***Model Curriculum for Data Science (MC-DS)*** as a template for building customisable Data Science curricula based on the proposed CF-DS and DS-BoK.

Objective 2: Education and Training Environment

Provide environment for ***re-skilling*** *and* ***certifying*** Data Scientists expertise to graduates, practitioners and researchers throughout their careers.

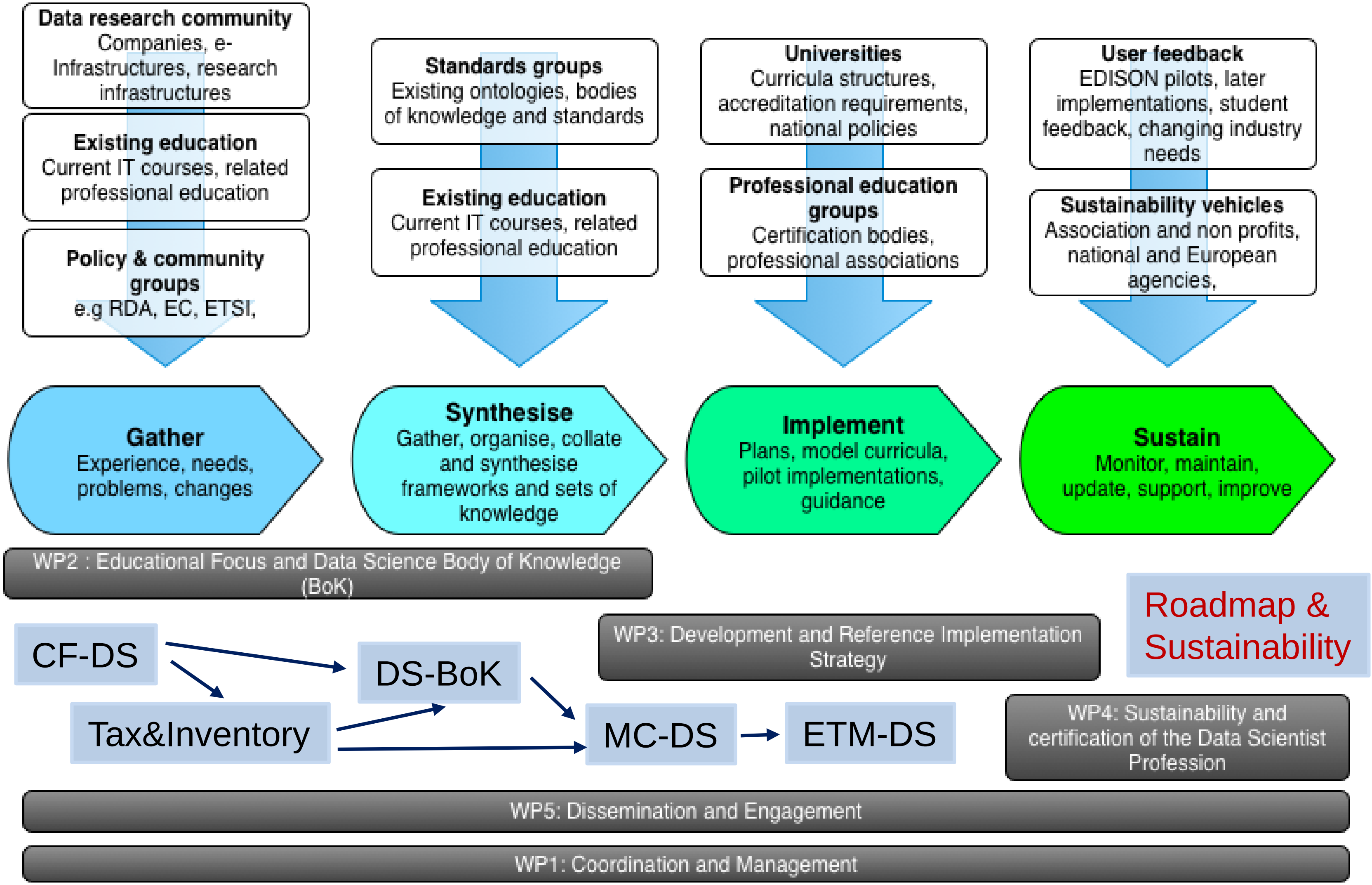
- Create EDISON **Education and Training Marketplace** by leveraging EGI Engage Training Marketplace.

Objective 3: Sustainability Model

Develop a ***sustainable business model*** and a ***roadmap*** for European education and training on Data Science technologies, provide a basis for the formal recognition of the Data Scientist as a new profession

- Establish networks and community of **champion universities**
- Create Community of practice for sustainable Data Science education and training supported by **EDISON Liaison Group(s)**

Basic methodology of EDISON: Development flow, work packages, and products



Data research community
Companies, e-Infrastructures, research infrastructures

Existing education
Current IT courses, related professional education

Policy & community groups
e.g RDA, EC, ETSI,

Standards groups
Existing ontologies, bodies of knowledge and standards

Existing education
Current IT courses, related professional education

Universities
Curricula structures, accreditation requirements, national policies

Professional education groups
Certification bodies, professional associations

User feedback
EDISON pilots, later implementations, student feedback, changing industry needs

Sustainability vehicles
Association and non profits, national and European agencies,

Gather
Experience, needs, problems, changes

Synthesise
Gather, organise, collate and synthesise frameworks and sets of knowledge

Implement
Plans, model curricula, pilot implementations, guidance

Sustain
Monitor, maintain, update, support, improve

WP2 : Educational Focus and Data Science Body of Knowledge (BoK)

WP3: Development and Reference Implementation Strategy

Roadmap & Sustainability

WP4: Sustainability and certification of the Data Scientist Profession

CF-DS

DS-BoK

Tax&Inventory

MC-DS

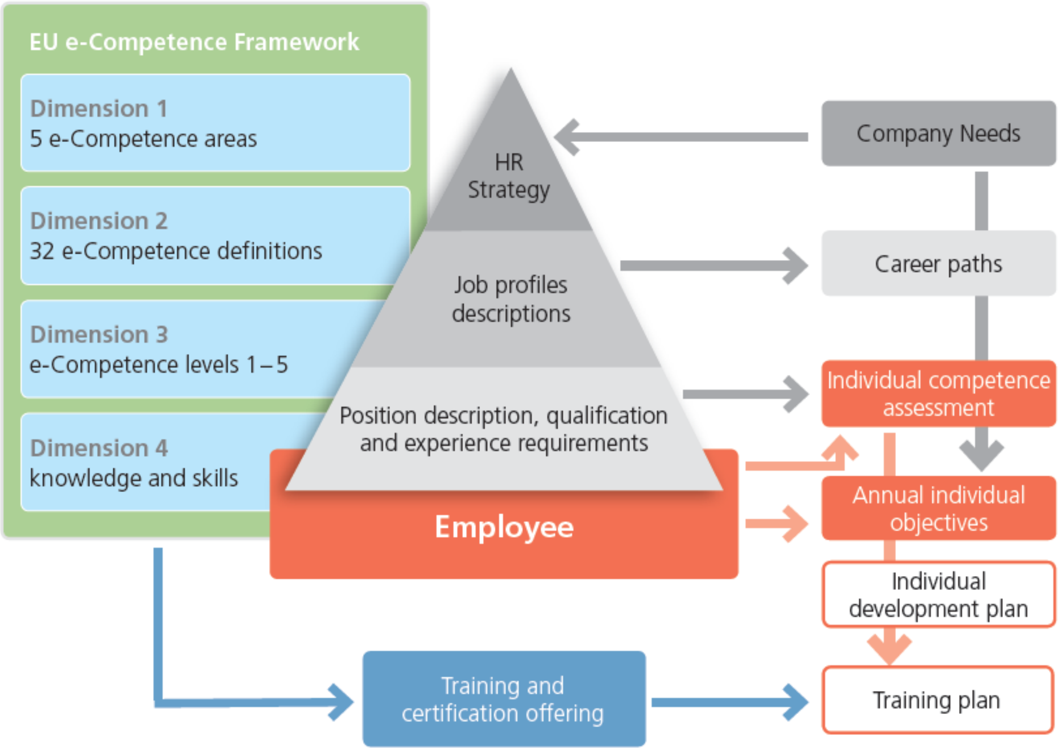
ETM-DS

WP5: Dissemination and Engagement

WP1: Coordination and Management

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 in a DS *Model Curriculum (MC-DS)*

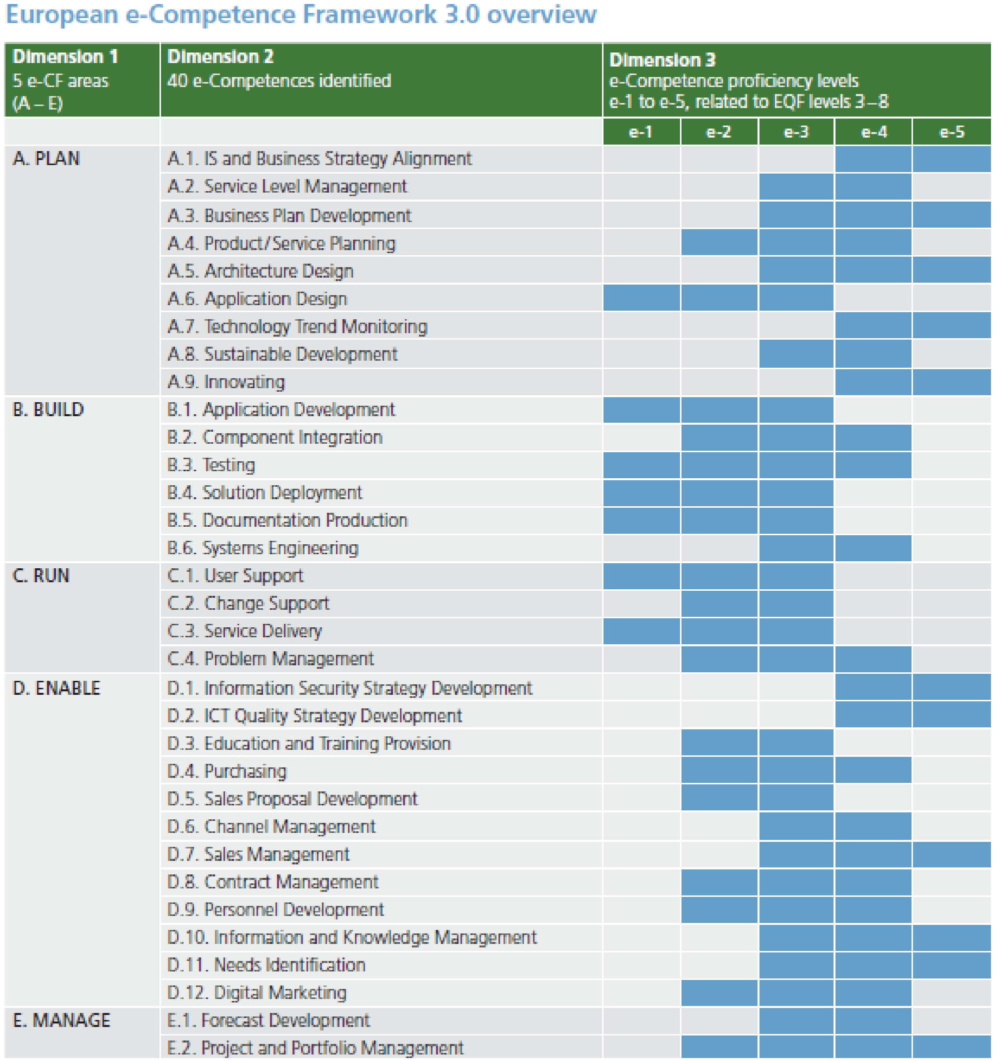


EU e-Competence Framework
Dimension 1
5 e-Competence areas
Dimension 2
32 e-Competence definitions
Dimension 3
e-Competence levels 1 – 5
Dimension 4
knowledge and skills
HR Strategy
Job profiles descriptions
Position description, qualification and experience requirements
Employee
Company Needs
Career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

- 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
  - 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 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)

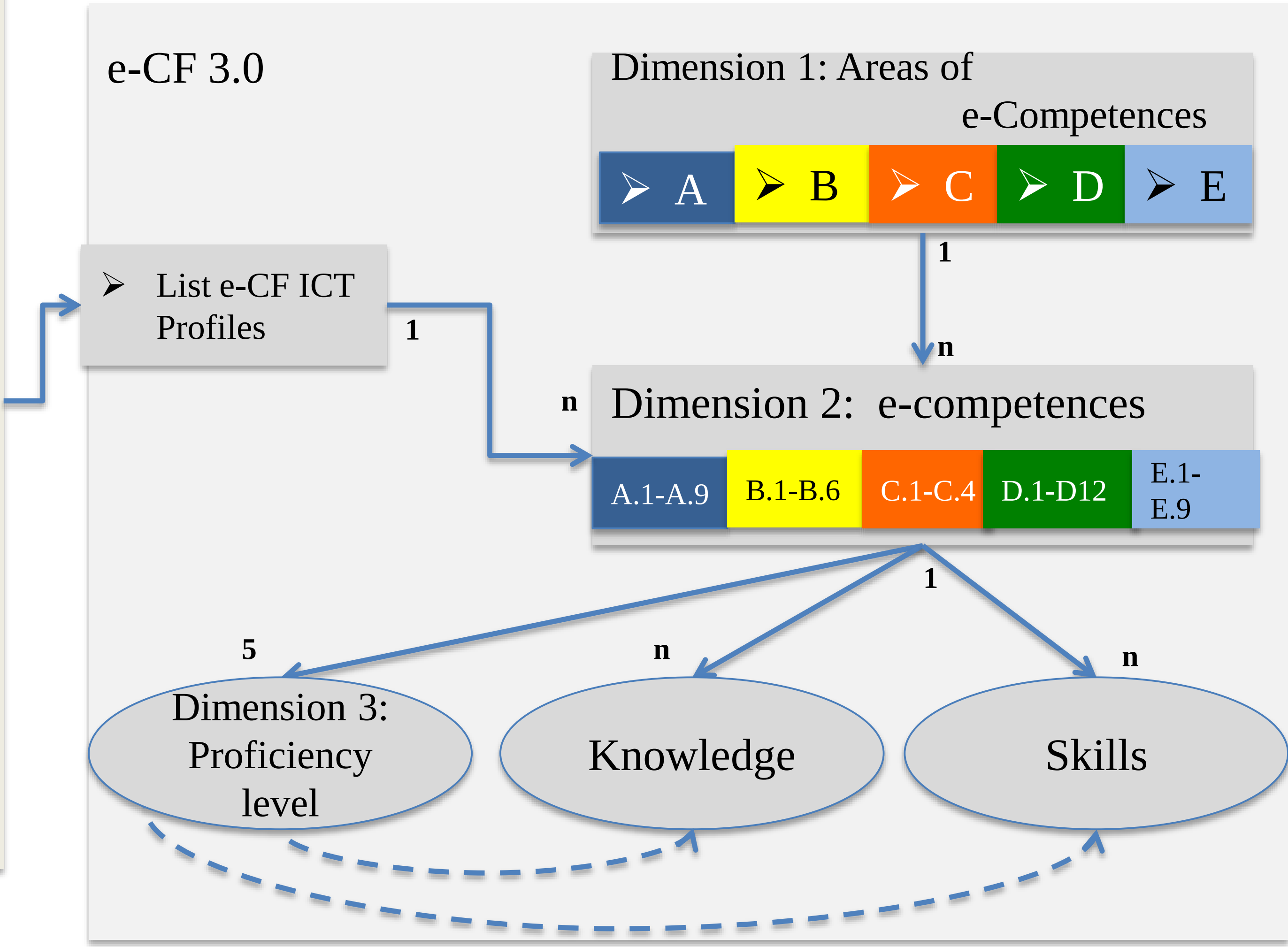
Definitions (according to e-CFv3.0)

- **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 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



e-CF 3.0

Dimension 1: Areas of e-Competences

➢ A | ➢ B | ➢ C | ➢ D | ➢ E

➢ List e-CF ICT Profiles

Dimension 2: e-competences

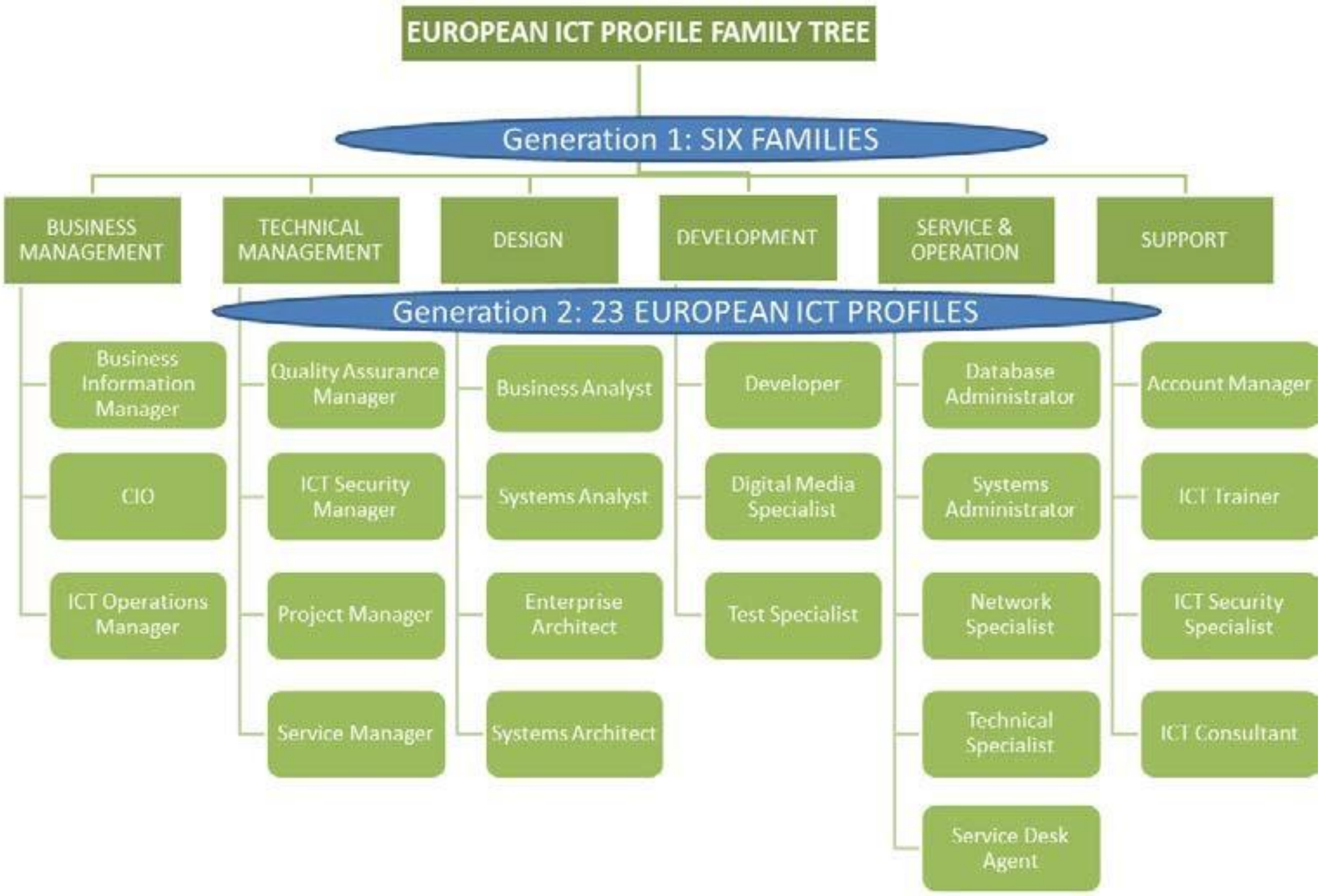
A.1-A.9 | B.1-B.6 | C.1-C.4 | D.1-D12 | E.1-E.9

Dimension 3: Proficiency level

Knowledge

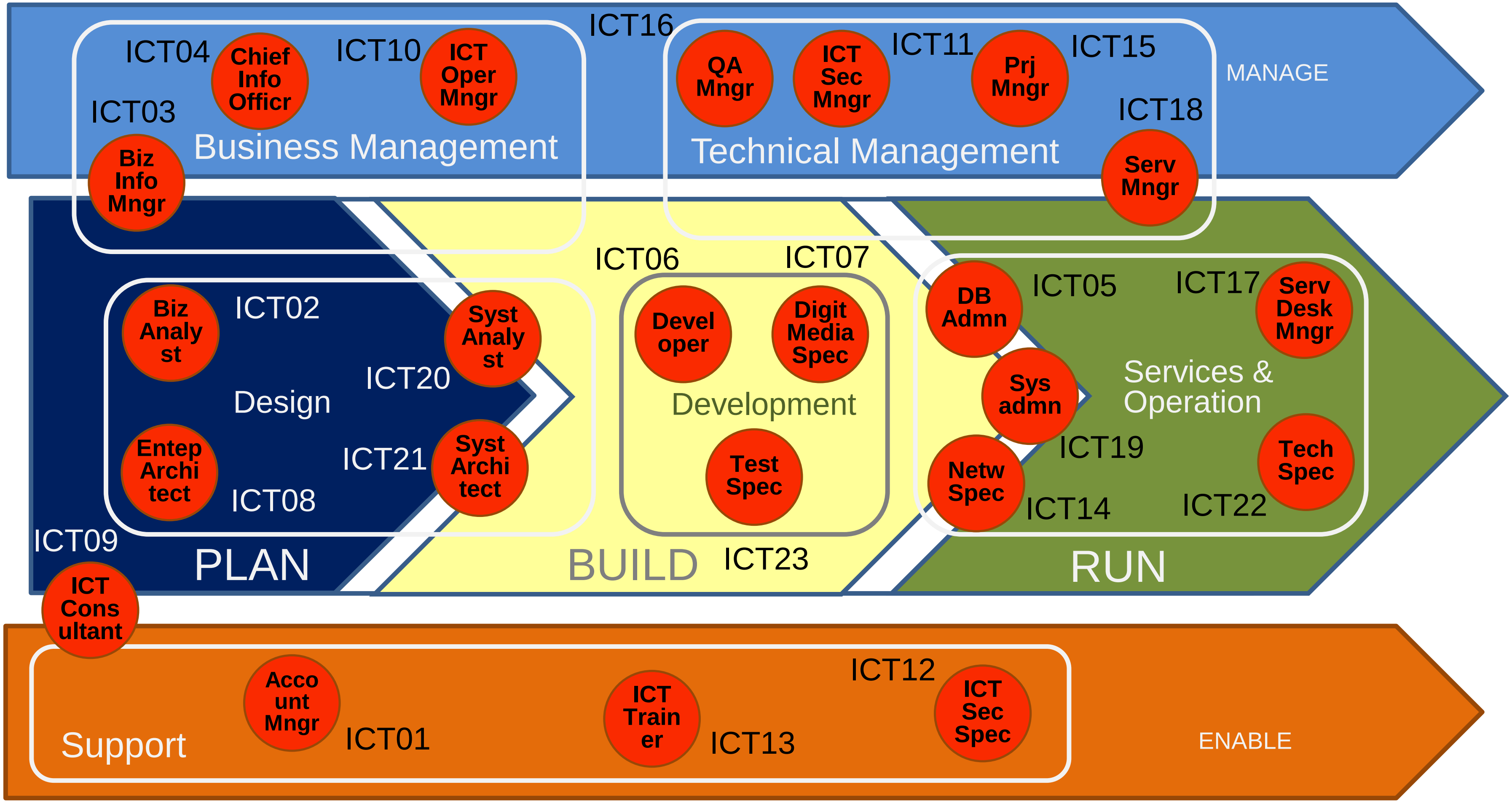
Skills

CWA 16458 (2012): European ICT Professional Profiles Family Tree



- European ICT Profile Family Tree – Generation 1 and 2 as a shared European reference

Mapping between e-CF3.0 and European ICT Profiles



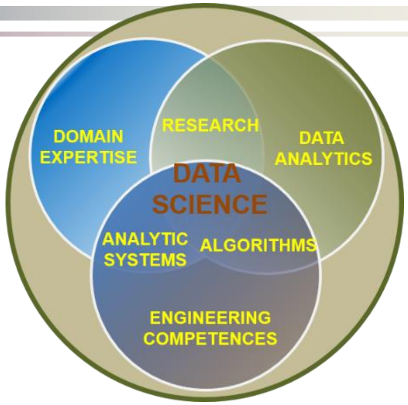
- European ICT Professional Profiles structured by six families and positioned within the ICT Business Process (e-CF Dimension 1)

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

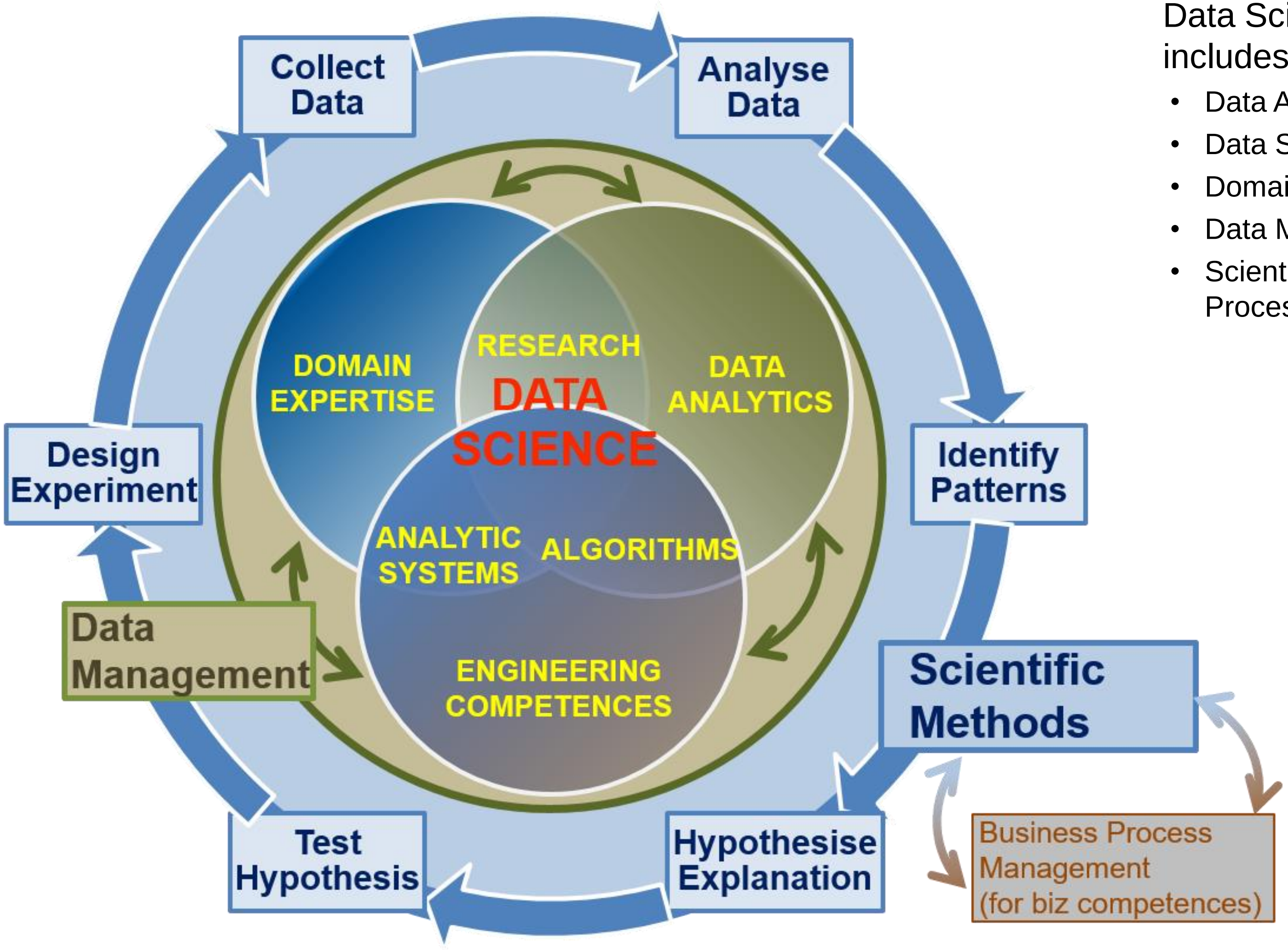
Identified Data Science Competence Groups

- 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: Provide such literacy advice and guiding to organisation*



[ref] Legacy: NIST BDWG definition of Data Science

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)

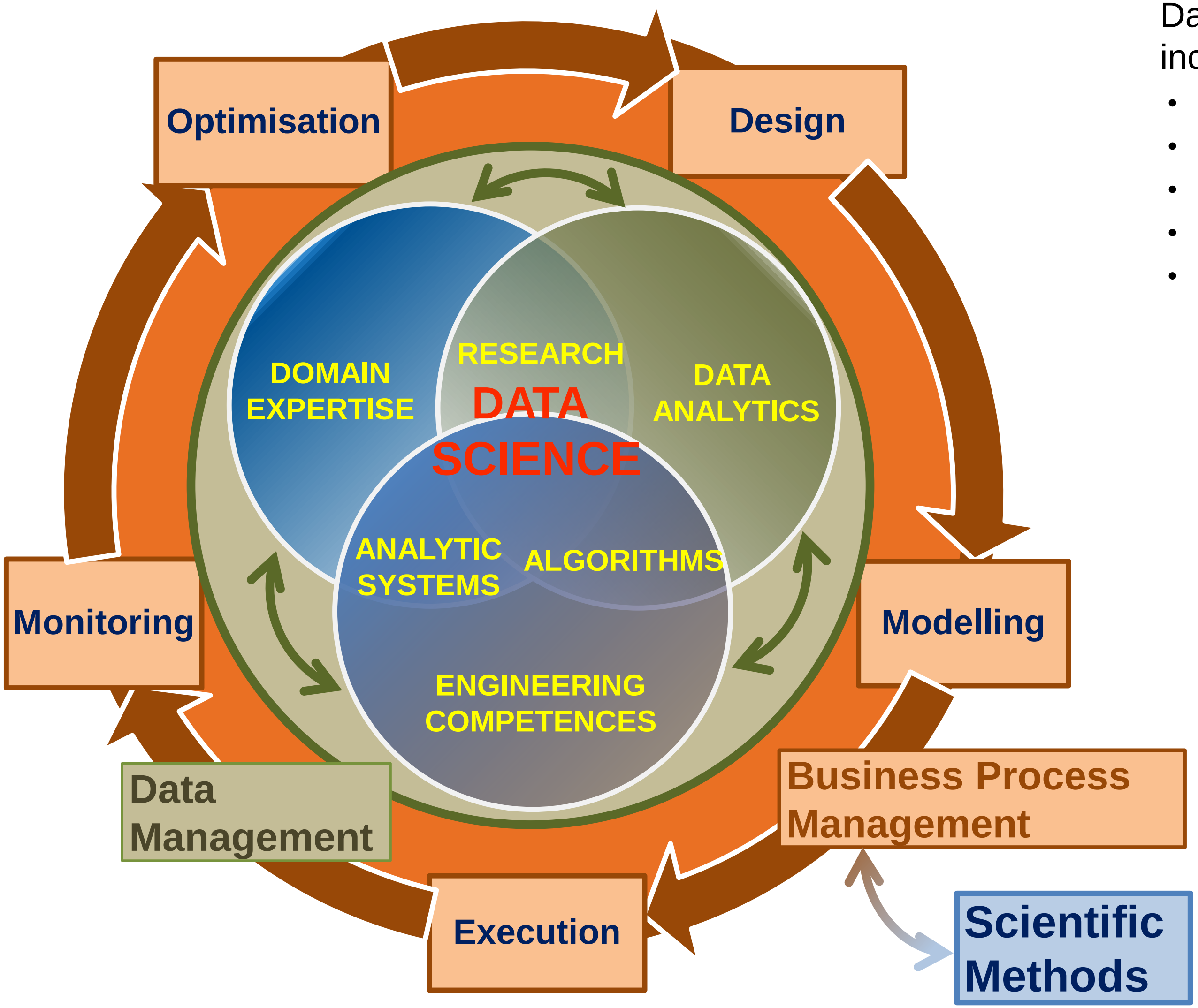
Scientific Methods

- 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 Areas – Business e-CF 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 Process Operations/Stages

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

Identified Data Science Skills/Experience Groups

- 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
  - Personal, inter-personal communication, team work (also called social intelligence or soft skills)
  - Application/subject domain related (research or business)
  - Mathematics and Statistics
- Big Data (Data Science) tools and platforms
  - Big Data Analytics platforms
  - Math& Stats tools
  - Databases (SQL and NoSQL)
  - Data Management and Curation platform
  - Data and applications visualisation
  - *No cloud related skills and knowledge mentioned explicitly*
- Programming and programming languages and IDE
  - General and specialized for data analysis and statistics
- Interpersonal skills (social intelligence)

Identified Data Science Skill Groups

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

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 (HortonWorks, Vertica LexisNexis HPCC System, etc)	Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)			

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)

Need for separate Data Science Competence Framework definition?

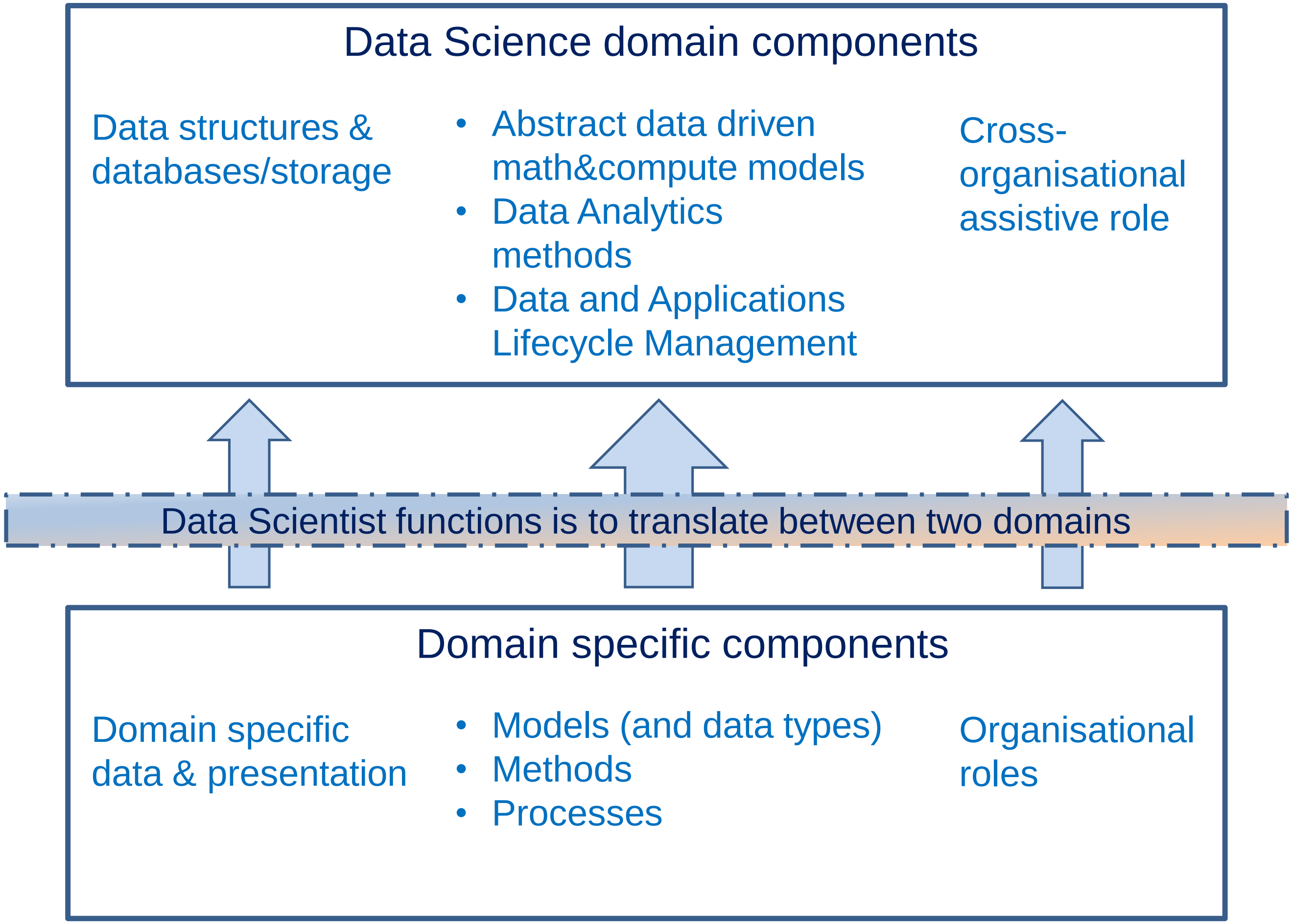
- There is no direct mapping of required/identified DS competences and skills to e-CF areas (i.e. organizational workflow stages)
  - Data Scientist is involved into all stages/areas
  - In most cases Data Scientist competences are connected to Data Lifecycle and not organizational workflow
- Data Scientist is a cross-intra-organizational role
  - Interact with different roles
  - Deliver information to top management
- Initially assistive but may play key (leading) role in data driven organizational processes and services
  - Potentially may have best organizational insight
  - Provide a basis for a future CEO mindset
- e-CF3.0 extensions with specific Data Science competences as a first step

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



Data Science domain components

Data structures & databases/storage

- 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

- Models (and data types)
- Methods
- Processes

Organisational roles

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

Mapping Data Scientist Competences to e-CF3.0

- Use e-CF to identify
  - ICT/CS background competences/skills/knowledge
  - Required basic education and training

- Use Data Science CF-DS to define additional training/re-skilling
  - From general ICT/(Librarian) to Data Scientist
    - Specialised Data Science courses
  - From Data Scientist practitioner to certified Data Scentist
    - General IT and CS courses

EXAMPLE: e-CF Dimensions for RI Technical (based on RDA IG-ETRD work)

- Dimension 1: 5 e-Competence areas, derived from the ICT processes present in RI development, management and operation:
  A. PLAN and DESIGN
  B. BUILD: DEVELOP and DEPLOY/IMPLEMENT
  C. OPERATE (RUN)
  D. USE: UTILISE (ENABLE)
  E. MANAGE

- Dimension 2: A set of reference competences for each area; currently identified 35 competences that are mapped from the general eCFv3.0.

- Dimension 3: Proficiency levels of each e-Competence, currently using eCF approach that provides European reference level specifications on e-Competence levels e-1 to e-5, which are related to the EQF levels 3 to 8.

- Dimension 4: Samples of knowledge and skills related to e-Competences in dimension 2. They will be provided to add value and context and are not intended to be exhaustive.

EXAMPLE: How to use eCF for New Profile of RI Technical

A. PLAN and DESIGN

A.2. Service Level Management
A.3. Product / Service Planning
A.5. Application Design
A.4. Architecture Design

Additional

A.6. Sustainable Development
A.7. Innovating and Technology Trend Monitoring
A.8. Business/Research Plan Development and Grant application
A.1. RI and Research Strategy Alignment

B. BUILD: DEVELOP and DEPLOY/IMPLEMENT

B.1. Application Development (Reqs Engineering, Function Specs, API, HCI)
B.2. Component Integration
B.3. Testing (RI services and Sci Apps)
B.4. Solution/Apps Deployment

Additional

B.5. Documentation Production
B.6. Systems Engineering (DevOps)

C. OPERATE (RUN)

C.1. User Support
C.2. Service Delivery
C.3. Problem Management

Additional

C.4. Change Support (Upgrade/Migration)

D. USE: UTILISE (ENABLE)

D.1. Scientific Applications Integration (on running RI)
D.5. Data collection and preservation
D.4. New requirements and change Identification
D.6. Education and Training Provision

Additional

D.2. Information Security Strategy Development
D.3. RI/ICT Quality Strategy Development
D.7. Purchasing/Procurement
D.8. Contract Management
D.9. Personnel Development
D.10. Dissemination and outreach

E. MANAGE

E.1. Overall RI management (by systems and components)
E.5. Information/Data Security Management

Additional

E.6. Data Management (including planning and lifecycle management, curation)
E.4. RI Security and Risk/Dependability Management
E.2. Project and Portfolio Management
E.3. ICT Quality Management and Compliance
E.7. RI/IS Governance

Further Steps

- Define a taxonomy and classification for DS competences and skills as a basis for more formal CF-DS definition
  - Closer look at skills, tools and platforms
- Suggest e-CF3.0 extensions and present them at ELG meeting and workshop in Bari
  - Provide feedback and contribution to CEN workshop on e-Competence for their meeting on 9 Dec 2015 in Paris
  - Talk to national e-CF bodies or adopters if available
- 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
      - Explore LinkedIn opportunities
  - Plan a number of key interviews, primarily experts and top executives at universities and companies
- Provide input to DS-BoK definition following from CF-DS
  - Link/Map to taxonomy of academic and educational and training courses
- Start related Social Network activity to promote already obtained results related to identified new Data Scientist competences and skills