EDISON
Data Science Competence Framework (CF-DS)
and
Data Science Body of Knowledge (DS-BoK)

Yuri Demchenko, EDISON
University of Amsterdam

Competences and Skills for Data and Research Infrastructures

Date: 29th February, 2016
RDA7, Tokyo, Japan

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
- Background: Existing frameworks and standards
  - e-CF3.0, CWA ICT profiles, ESCO
- Data Science Competence Framework: Essential competences and skills
  - Domain related competences and skills
  - Taxonomy: Occupations and competences
- Organisational processes and role of Data Scientist
- Data Science Body of Knowledge (DS-BoK)
  - Taxonomy: Knowledge area, academic disciplines
- Further steps - Survey and questionnaires

UvA Universiteit van Amsterdam

University of Southampton

inmark europa

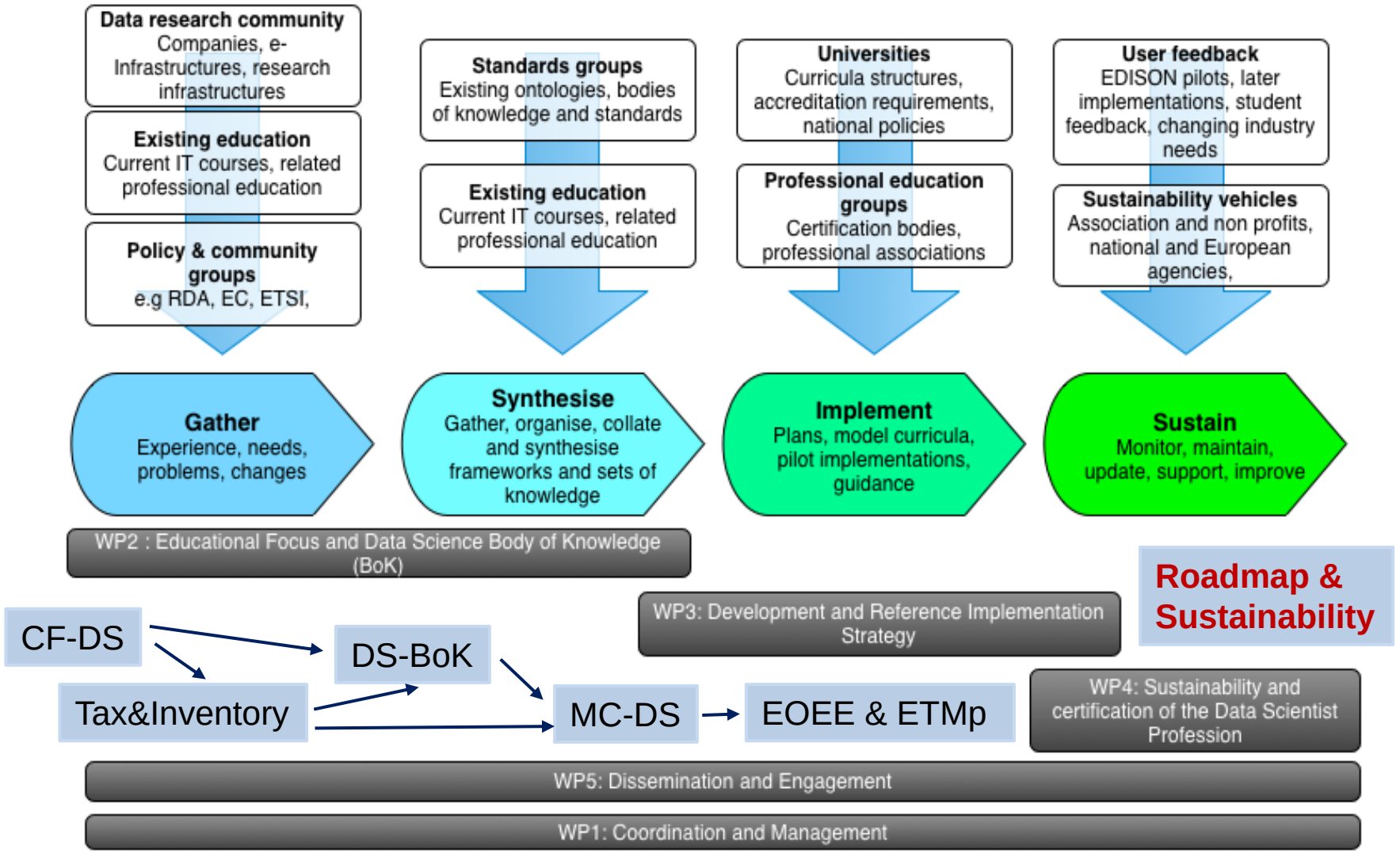
University of Stavanger

EGI

ENGINEERING

FTK Forschungsinstitut für Telekommunikation und Kooperation

EDISON methodology: Development flow, work packages, and products

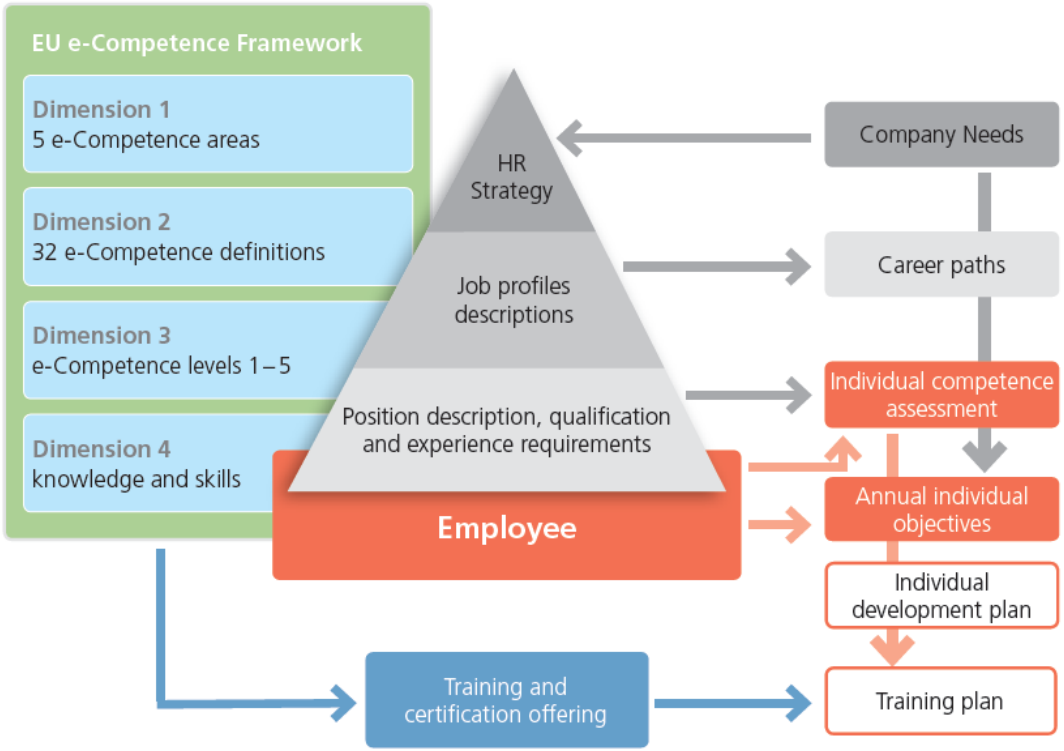


Background: EU Competence Frameworks and Profiles

- e-CFv3.0 - European e-Competence framework for IT
- CWA 16458 (2012): European ICT Professional Profiles Family Tree
- ESCO (European Skills, Competences, Qualifications and Occupations) framework

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 cycle/flow*, 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)*



- 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

-> Refactor to Scientific Research cycle/workflow (and linked to 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 can be treated as outcome of learning or training

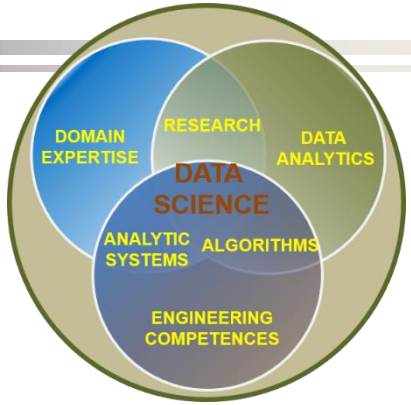
- **Knowledge** in the context of competence definition is treated as something to know, to be aware of, familiar with, and obtained as a part of education.

- **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

- Sources (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
- 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 are without specified skills profile
  - Explicit Data Scientist jobs specify wide variety of expected functions/responsibilities and required skills and knowledge

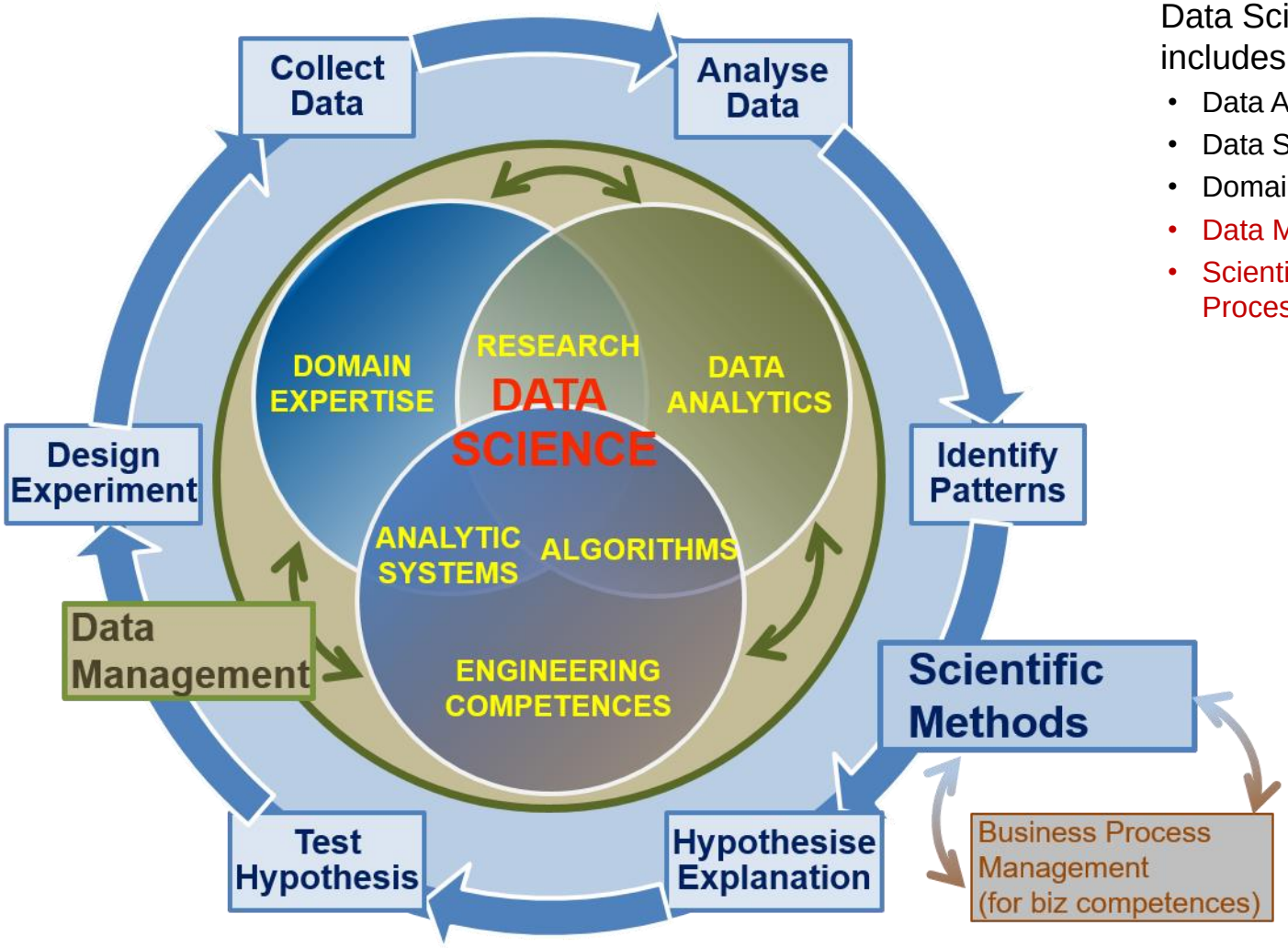
Identified Data Science Competence Groups



- Traditional/known Data Science competences/skills groups 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 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
  - 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)

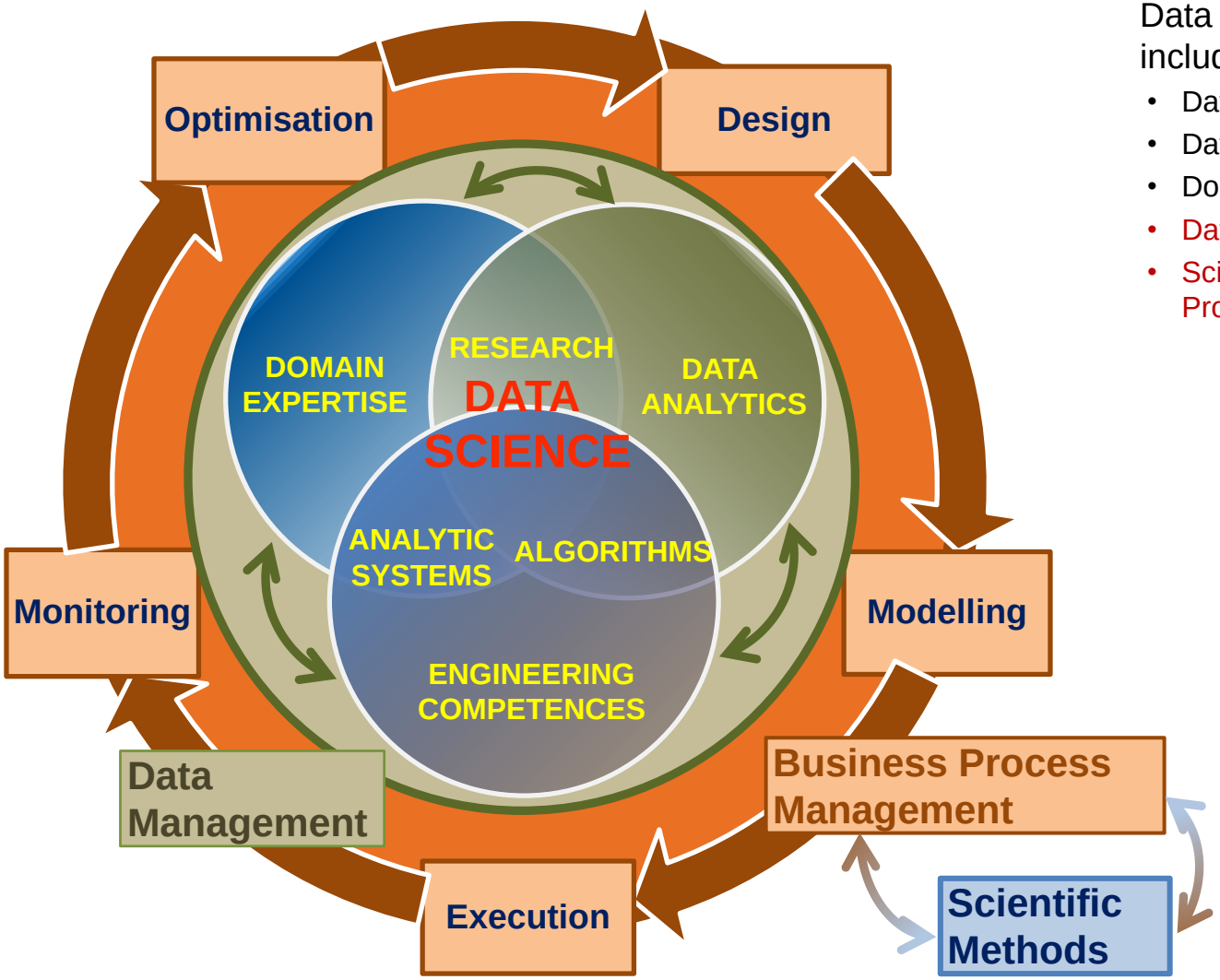
Scientific Methods

- 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 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

	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
  - *Cloud based platforms and tools*
- Programming and programming languages and IDE
  - General and specialized for data analysis and statistics

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)			

- Big Data Analytics platforms
- Math& Stats tools
- Databases
- Data/applications visualization
- Data Management and Curation platform

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

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

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)

Possible Data Scientist profiles/roles as extension to CWA16458 (2012) or ESCO

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

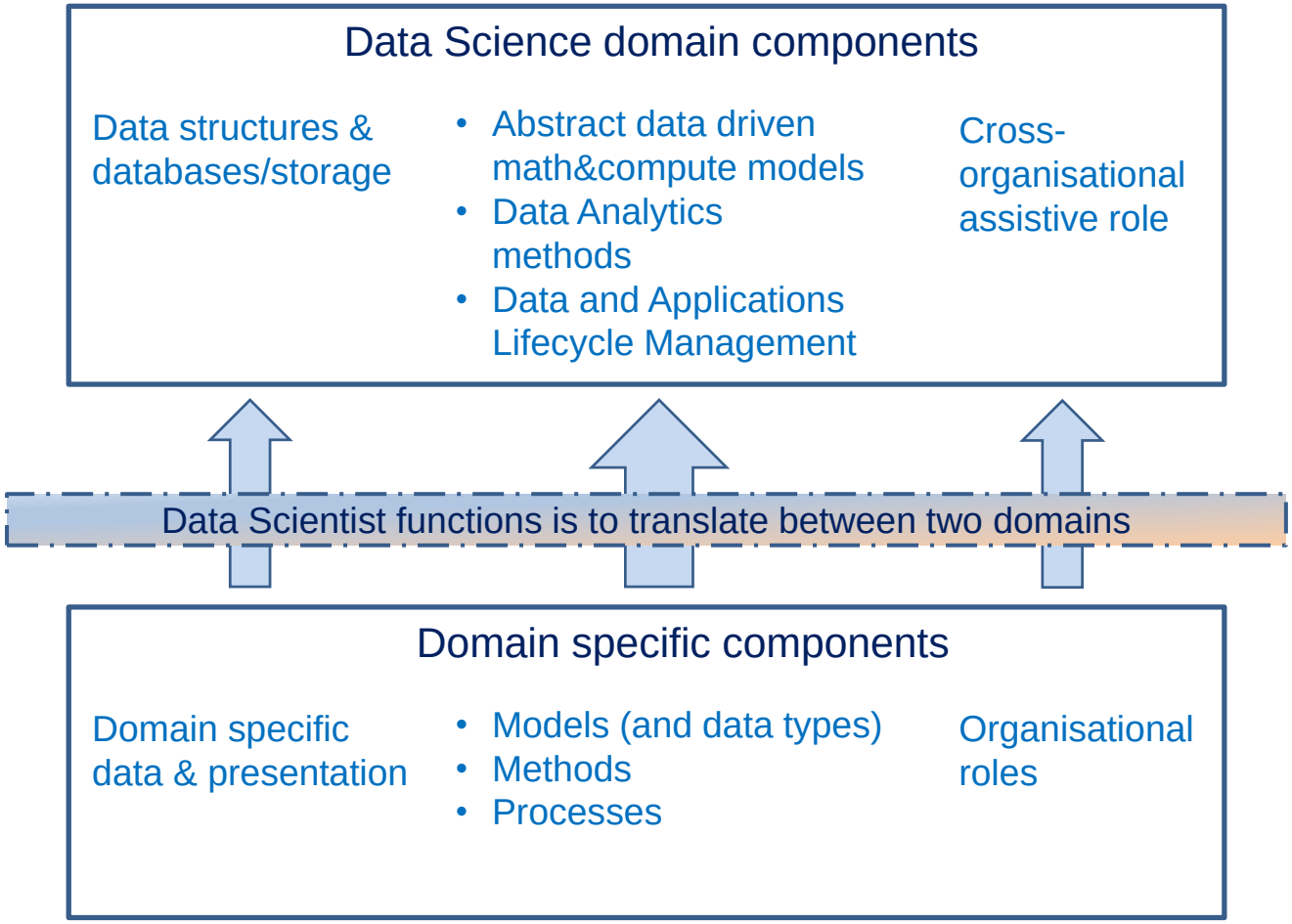
- Research e-Infrastructure brings its own specifics to required competences and skills definition

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



EXAMPLE: Use of e-CF3.0 for Defining Profile of RI Technical (part of RDA IG-ETRD work)

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 Scientific 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

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 version2 “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, ORCID

(13) Data Management Plan

(14) Research Data Infrastructure

European consultation call – Deadline 30 April 2016

https://ec.europa.eu/futurium/en/content/consultation-european-e-infrastructure

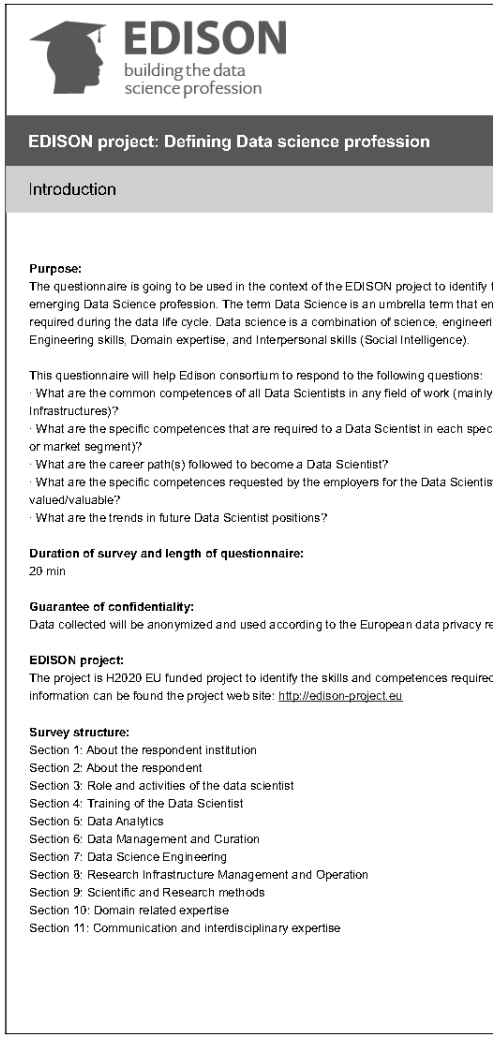
- What are the main challenges for the realisation of an integrated European e- infrastructure from the perspective of scientific data-related needs (from data access to sharing, analytics, re-use, preservation, standards, interoperability, value chain and other issues)?
- What are the challenges for reinforcing the cooperation between European e-infrastructure service providers and their scientific users, including thematic research infrastructures, to accelerate user's adoption of e-infrastructure services - such as identity management innovation - and foster innovation in e-infrastructures?
- What are the challenges faced by industrial actors preventing them to fully benefit from the services provided by European e-infrastructures and to contribute to the innovation of the existing e-infrastructures?
- **What are the main challenges Europe is facing regarding skills and competences required for effective data driven science, and management of research e-infrastructures?**

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
- Create a Questionnaire and run Survey using CF-DS vocabulary
  - 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
- Proceed with suggested e-CF3.0 extensions and participate in the next e-CF meetings
  - Talk to national e-CF bodies or adopters if available
- Provide feedback and contribution to ESCO
- Suggest ACM2012 Classification extensions and contact ACM people
- Provide input to DS-BoK definition following from CF-DS
  - Link/Map to taxonomy of academic and educational and training courses
- Create open community forum to collect contribution
  - CF-DS document is on public comments available from EDISON website http://www.edison-project.eu/data-science-competence-framework-cf-ds
  - Start related Social Network groups to promote already obtained results and obtain feedback and community contribution

Survey

Survey link https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession



EDISON
building the data science profession

EDISON project: Defining Data science profession

Introduction

Purpose:
The questionnaire is going to be used in the context of the EDISON project to identify t
emerging Data Science profession. The term Data Science is an umbrella term that en
required during the data life cycle. Data science is a combination of science, engineeri
Engineering skills, Domain expertise, and Interpersonal skills (Social Intelligence).

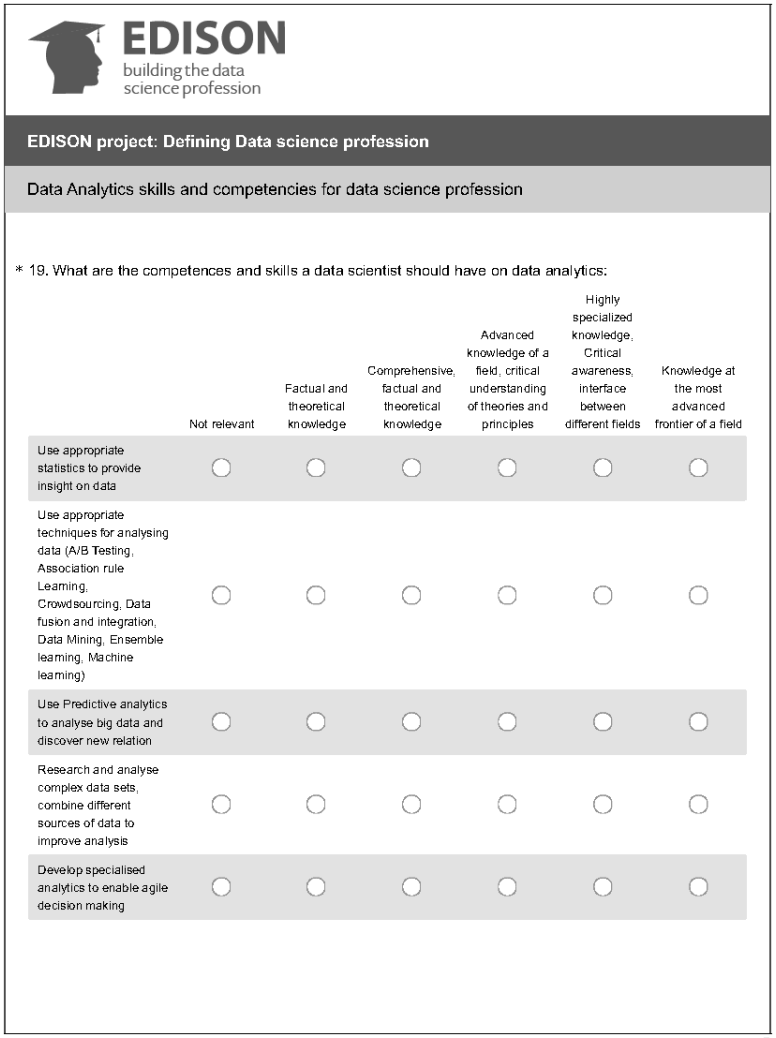
This questionnaire will help Edison consortium to respond to the following questions:
· What are the common competences of all Data Scientists in any field of work (mainly
Infrastructures)?
· What are the specific competences that are required to a Data Scientist in each spec
or market segment)?
· What are the career path(s) followed to become a Data Scientist?
· What are the specific competences requested by the employers for the Data Scientis
valued/valuable?
· What are the trends in future Data Scientist positions?

Duration of survey and length of questionnaire:
20 min

Guarantee of confidentiality:
Data collected will be anonymized and used according to the European data privacy re

EDISON project:
The project is H2020 EU funded project to identify the skills and competences requirec
information can be found the project web site: http://edison-project.eu

Survey structure:
Section 1: About the respondent institution
Section 2: About the respondent
Section 3: Role and activities of the data scientist
Section 4: Training of the Data Scientist
Section 5: Data Analytics
Section 6: Data Management and Curation
Section 7: Data Science Engineering
Section 8: Research Infrastructure Management and Operation
Section 9: Scientific and Research methods
Section 10: Domain related expertise
Section 11: Communication and interdisciplinary expertise



EDISON
building the data science profession

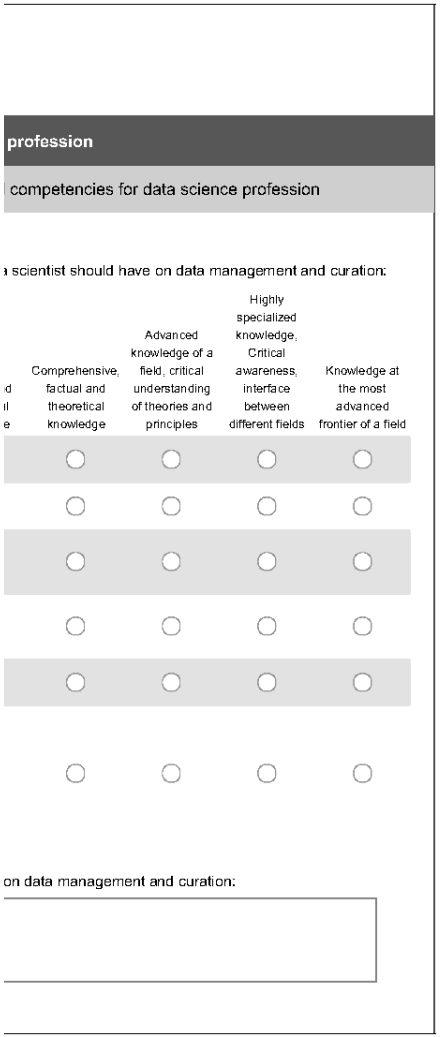
EDISON project: Defining Data science profession

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



profession

competencies for data science profession

a scientist should have on data management and curation:

	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
d	○	○	○	○
il	○	○	○	○
e	○	○	○	○
	○	○	○	○
	○	○	○	○
	○	○	○	○

on data management and curation:

10