

EDISON Services for Core Data Expert Capacity Building and Skills Management

EDISON Project update



EDISON – Education for Data Intensive Science to Open New science frontiers

Grant 675419 (INFRASUPP-4-2015: CSA)

Yuri Demchenko, EDISON Project University of Amsterdam

RDA 9 IG-ETHRD meeting 5 April 2017, Barcelona

EDISON **EDISON Services Overview: Capacity Building and Skills Management** science profession

EDISON Data Science Framework (EDSF)

building the data

- Compliant with EU standards on competences and professional occupations e-CFv3.0, ESCO
- Customisable courses design for targeted education and training
- Skills development and career management for Core Data Experts and related data handling professions
- Capacity building and Data Science team design
- Academic programmes and professional training courses (self) assessment and design
- EU network of Champion universities pioneering Data Science academic programmes
- Engagement in relevant RDA activities and groups
- Cooperation with International professional organisations IEEE, ACM, BHEF, APEC (AP Economic Cooperation)







EDISON Data Science Framework (EDSF) Release 1 (October 2016)



- EDISON Framework components
 - CF-DS Data Science Competence Framework
 - DS-BoK Data Science Body of Knowledge
 - MC-DS Data Science Model Curriculum
 - DSP Data Science Professional profiles
 - Data Science Taxonomies and Scientific Disciplines Classification
 - EOEE EDISON Online Education Environment

2017

Outcome Based Educations and Training Model



From Competences and DSP Profiles to Learning Outcomes (LO) and to Knowledge Unites (KU) and Learning Units (LU)

• EDSF allow for customized educational courses and training modules design

DSP Profiles mapping to ESCO Taxonomy **High Level Groups**



DSP Profiles mapping to corresponding CF-DS Competence Groups Relevance level from 5 – maximum to 1 – minimum



Individual Competences Benchmarking



Individual Education/Training Path based on Competence benchmarking

- Red polygon indicates the chosen professional profile: Data Scientist (general)
- Green polygon indicates the candidate or practitioner competences/skills profile
- Insufficient competences (gaps) are highlighted in *red*
 - DSDA01 DSDA06 Data Science Analytics
 - DSRM01 DSRM05 Data Science Research Methods
- Can be use for team skills match marking and organisational skills management

[ref] For DSP Profiles definition and for enumerated competences refer to EDSF documents CF-DS and DSP Profiles.

Building a Data Science Team





Recent events and developments

• EDISON Champions conference 15-16 March 2017

- Data Steward role and profile definition
- Growth of librarians mastering Data Science
- Number of initiatives -> Madrid Communique to be published
- PwC and BEHF report "Investing in America's data science and analytics talent: The case for action"
 - Published 30 March 2017
 - Contributed by EDISON partners
- **DARE project by APEC** (Asia Pacific Economic Cooperation)
 - Initial: Development of the Data Science Analytics checklist
 - Extend to Body of Knowledge, professional profiles and skills management
- IEEE/ACM and other professional communities in USA
 - 4 March 2017 Meeting of Professional and Academic Societies to Explore Feasibility of a Joint Project for Developing Curriculum Recommendation(s) for Data Science



×

Data Scientist definition

Based on the definitions by NIST Big Data WG (NIST SP1500 - 2015)

- A Data Scientist is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle
 - Till the delivery of expected scientific and business value to science or industry



- Ability to solve variety of business problems
- Optimize performance and suggest new services for the organisation
- Develop a special mindset and be statistically minded, understand raw data and "appreciate data as a first class product"
- **Data science** is the empirical synthesis of actionable knowledge and technologies required to handle data from raw data through the complete data lifecycle process.
- **Big Data** is the technology to build system and infrastructures to process large volume of structurally complex data in a time effective way



[ref] Legacy: NIST BDWG definition of Data Science

Data Science Competence Groups - Research



Data Science Competences Groups – Business



RDA9 IG-ETHRD - 5 April 2017

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 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



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)



- EDISON project website <u>http://edison-project.eu/</u>
- EDISON Data Science Framework Release 1 (EDSF)
 <u>http://edison-project.eu/edison-data-science-framework-edsf</u>
 - Data Science Competence Framework
 <u>http://edison-project.eu/data-science-competence-framework-cf-ds</u>
 - Data Science Body of Knowledge
 <u>http://edison-project.eu/data-science-body-knowledge-ds-bok</u>
 - Data Science Model Curriculum
 <u>http://edison-project.eu/data-science-model-curriculum-mc-ds</u>
 - Data Science Professional Profiles
 <u>http://edison-project.eu/data-science-professional-profiles-definition-dsp</u>
- Survey Data Science Competences: Invitation to participate
 <u>https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession</u>

Data Scientist and Subject Domain Specialist

Subject domain components

- Model (and data types)
- Methods
- Processes
- Domain specific data and presentation/visualization methods
- Organisational roles and relations

Data Scientist is an assistant to Subject Domain Specialists

- Translate subject domain Model, Methods, Processes into abstract data driven form
- Implement computational models in software, build required infrastructure and tools
- Do (computational) analytic work and present it in a form understandable to subject domain
- Discover new relations originated from data analysis and advice subject domain specialist
- Present/visualise information in domain related actionable way
- Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data

Data Science and Subject Domains



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

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