EDISON Data Science Framework (EDSF) as a foundation for the Data Science education, training and sustainable skills development

Yuri Demchenko, EDISON Project
University of Amsterdam

e-IRG Workshop
8-9 June 2017, Malta
Outline

• Background
  – Recent EU Initiatives, European Digital Single Market (DSM) and demand for data enabled skills
• EDISON Data Science Framework (EDSF)
  – From Data Science Competences and Skills to Body of Knowledge and Model Curriculum
  – Data Science Profession Profiles family and organisational skills management
• Use of EDSF for Data Science curricula design and organisational skills management
• Summary and discussion

- The need for new multidisciplinary and digital skills in particular Data Scientist
  - Expected rapidly growing demand will lead to more than 800 000 unfilled vacancies by 2020


- Launch Digital Skills and Jobs Coalition (1st December 2016, Brussels) to develop comprehensive national digital skills strategies by mid-2017


- 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

HLEG report on European Open Science Cloud (October 2016) identified need for data experts and data stewards

- Estimation: More than 80,000 data stewards (1 per every 20 scientists)
- Core Data Experts need to be trained and their career perspective improved
Industry report on Data Science Analytics and Data enabled skills demand

  - 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%

- PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017)
  - 2.5 mln postings, 23% Data Scientist, 67% DSA enabled jobs
  - DSA enabled jobs growing at higher rate than main Data Science jobs

- Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017)
  - DSA enabled jobs takes 45-58 days to fill: 5 days longer than average
ICT and Data related Skills shortage (in life)

• Problems with hiring (skilled) ICT related staff
  – At least one year for training and acquiring experience
  – As soon as new employees are confident with their skills, they leave for big companies or industry

• Open Data Science/Stewards positions stay unfilled
  – In research institutions for months and years
  – In industry for months

• Companies/organisations want experienced Data Science workers
  – There is no time to acquire necessary experience

• Millennials factor
  – Do we understand difference of the millennials workforce?

• Challenge: How to obtain, train and sustain new digital (ICT and Data related) skills in organisations
Sustainable ICT and Data Skills Development

• HLEG report on EOSC rose question about critical need for Core Data Experts
  – Any changes since report publication in October 2016?

• Educate vs Train
  – Training is a short term solution
  – Education is a basis for sustainable skills development

• Technology focus changes every 3-4 years
  – Study: 50% of academic curricula are outdated at the time of graduation

• Lack of necessary skills leads to underperforming projects and organisations and lose of competitiveness

• Need to change skills management paradigm
  – Dynamic (self-) re-skilling: Continuous professional development and shared responsibility between employer and employee
  – Skills and career management as a part of professional orientation
• EDISON Data Science Framework (EDSF)
  – 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)

**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

**Methodology**
- ESDF development based on job market study, existing practices in academic, research and industry.
- Review and feedback from the ELG, expert community, domain experts.
- Input from the champion universities and community of practice.
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 an **expected scientific and business value** to science or industry

- **Profession is defined via Competences mapped to**
  - **Skills and Knowledge**
  - **Proficiency levels**

- **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
Data Science Competence Groups - Research

Data Science Competences include 5 groups:
- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and 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 Competences include 5 groups
- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and 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
<table>
<thead>
<tr>
<th>Identified Data Science Competence Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Science Analytics (DSDA)</strong></td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>
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 21st Century Skills**
  - Personal, inter-personal communication, team work, professional network
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
  - For customizable training and career development
  - Including CV or organisational profiles matching

- 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
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

Malta, 8-9 June 2017
**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 (data handling/management):** Data Stewards, Digital Data Curator, Digital Librarians, Data Archivists

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

**Support workers and data handling clerks:** User support workers, data entry clerks, data entry field workers

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

• Foundation and methodological base
  – Data Science Body of Knowledge (DS-BoK)
    • Taxonomy and classification of Data Science related scientific subjects
  – Data Science Model Curriculum (MC-DS)
    • Set Learning Units mapped to CF-DS Learning and DS-BoK Knowledge Areas/Units
  – Instructional methodologies and teaching models

• Platforms and environment
  – Virtual labs, datasets, developments platforms
  – Online education environment and courses management

• Services
  – Individual benchmarking and profiling tools (competence assessment)
  – Knowledge evaluation tools
  – Certifications and training for self-made Data Scientists practitioners
  – Education and training marketplace: Courses catalog and repository
Outcome Based Educations and Training Model

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

- EDSF allow for customized educational courses and training modules design
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

Data Collection

Data Ingest

Data Analysis

Results, Actionable Data

Data Source (Experiment, Data Driven Application)

Data Science Group Manager, Data Science Architect

Data Science Applications Developer

Data Steward

Data Engineer, Database Developer

Data Analyst/Business Analyst

Data Science Applications Developer

Data Steward

Data Scientist

Data Steward

Data Scientist

Data Steward

Data Science Researcher (Scientific domain)

Data Facilities Operator

Data Entry/Support

Data Steward

Data Science Researcher Business Analyst

Data Scientist

EDISON Data Science Framework (EDSF)
Data Stewards – A rising new role in Data Science ecosystem

• Data Stewards as a key bridging role between Data Scientists as (hard)core data experts and scientific domain researchers (HLEG EOSC report)

• Current definition of Data Steward (part of Data Science Professional profiles)
  – Data Steward is a data handling and management professional whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation. Data Steward creates data model for domain specific data, support and advice domain scientists/ researchers during the whole research cycle and data management lifecycle.
EDSF Recognition, Endorsement and Implementation

- **DARE (Data Analytics Rising Employment)** project by APEC (Asia Pacific Economic Cooperation)
  - DARE project Advisory Council meeting 4-5 May 2017, Singapore

- **PcW and BHEF Report “Investing in America’s data science and analytics talent”** April 2017
  - Quotes EDSF and Amsterdam School of Data Science

- **Dutch Ministry of Education recommended EDSF** as a basis for university curricula on Data Science
  - Workshop “Be Prepared for Big Data in the Cloud: Dutch Initiatives for personalized medicine and health research & toward a national action programme for data science training”, Amsterdam 28 June 2016

- **European Champion Universities network**
  - 1st Conference (13-14 July, UK), 2nd Conference (14-15 March, Madrid, Spain)
  - 3rd Conference 19-20 June 2017, Warsaw
Summary: Services and References

- EDISON Website [http://edison-project.eu/](http://edison-project.eu/)
- Directory of University programs [http://edison-project.eu/university-programs-list](http://edison-project.eu/university-programs-list)

- Survey Data Science Competences: Invitation to participate [https://www.surveymonkey.com/r/EDISON_project_Defining_Data_science_profession](https://www.surveymonkey.com/r/EDISON_project_Defining_Data_science_profession)

- Competences benchmarking and tailored training for practitioners
- Data Science Curriculum advice and design for universities
- Data Science team building and organizational roles profiling
Links to EDISON Resources

- **EDISON project website** [http://edison-project.eu/](http://edison-project.eu/)

- **EDISON Data Science Framework Release 1 (EDSF)**
  - Data Science Competence Framework
  - Data Science Body of Knowledge
  - Data Science Model Curriculum
  - Data Science Professional Profiles

- **Survey Data Science Competences: Invitation to participate**
  [https://www.surveymonkey.com/r/EDISON_project - Defining_Data_science_profession](https://www.surveymonkey.com/r/EDISON_project - Defining_Data_science_profession)
Other related links

• Amsterdam School of Data Science
  – https://www.schoolofdatascience.amsterdam/
  – https://www.schoolofdatascience.amsterdam/education/

• Research Data Alliance interest Group on Education and Training on Handling of Research Data (IG-ETHRD)
  – https://www.rd-alliance.org/groups/education-and-training-handling-research-data.html

• PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017)

• Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017)
Identified Data Science Competence Groups

• Core Data Science competences/skills groups
  – Data Science Analytics (including Statistical Analysis, Machine Learning, Business Analytics)
  – Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
  – Domain Knowledge and Expertise (Subject/Scientific domain related)

• EDISON identified 5 core 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 “21st Century Skills”
  – Inter-personal skills and team work, cooperativeness

• Important aspect of integrating Data Scientist (team) into organisation structure
  – General Data Science (and Data) literacy for all involved roles and management
  – Role of Data Scientist: Provide a kind of literacy advice and guidance to organisation
1. **Critical Thinking:** Demonstrating the ability to apply critical thinking skills to solve problems and make effective decisions
2. **Communication:** Understanding and communicating ideas
3. **Collaboration:** Working with other, appreciation of multicultural difference
4. **Creativity and Attitude:** Deliver high quality work and focus on final result, initiative, intellectual risk
5. **Planning & Organizing:** Planning and prioritizing work to manage time effectively and accomplish assigned tasks
6. **Business Fundamentals:** Having fundamental knowledge of the organization and the industry
7. **Customer Focus:** Actively look for ways to identify market demands and meet customer or client needs
8. **Working with Tools & Technology:** Selecting, using, and maintaining tools and technology to facilitate work activity
9. **Dynamic (self-) re-skilling:** Continuously monitor individual knowledge and skills as shared responsibility between employer and employee, ability to adopt to changes
10. **Professional networking:** Involvement and contribution to professional network activities
11. **Ethics:** Adhere to high ethical and professional norms, responsible use of power data driven technologies, avoid and disregard un-ethical use of technologies and biased data collection and presentation
Further developments and Next steps (1)

• Next EDSF release 2 (planned for June 2017) will link competences to skills and knowledge

• Final EDSF project deliverables (due August 2017) will include:
  – Data Science Education Sustainability Roadmap
    • Will involve wide consultation with experts community and also with EU policy makers
    • Will be reviewed by the EDISON Liaisons Groups (ELG)
  – Certification Framework for at least two levels of Data Science competences proficiency
    • Consultation with few certification providers is in the progress

• Toward EDSF and Data Science profession standardisation
  – ESCO (European Skills, Competences and Occupations) taxonomy – extending with the Data Science related occupations, competences and skills
  – CEN TC428 (European std body) – Extending current eCFv3.0 and ICT profiles towards e-CF4 with Data Science related competences
  – Work with the IEEE and ACM curriculum workshop to define Data Science Curriculum and extend current CCS2012 (Classification Computer Science 2012)

• Number of Case studies is planned in cooperation with active EU projects EDSA, EOSCpilot, BDVe, etc. (not limited to the project lifetime)
Further developments and Next steps (2)

- The EDISON project legacy will include (linked to the current project website and migrated to CP in the future)
  - EDSF – EDISON Data Science Framework
  - Data Science Community Portal (CP) - http://datasciencepro.eu/
  - EDISON project network including
    - EDISON Liaison Groups
    - Data Science Champions conference
    - Cooperative networks with European Research Infrastructures (e.g. HEP, Bioinformatics, Environment and Biodiversity, Maritime, etc),
      - International cooperative links BHEF, APEC, IEEE, ACM

- Applications and tools development
  - Prototypes will be produced in the timeline of the project but further development is a subject to additional funding

- Sustainability of the project legacy/products will be ensured by the project partners voluntarily for the period at least 3 years
  - EDSF will be maintained by UvA
  - CP by Engineering (Italy)
Further developments and Next steps (3)

- Further dissemination, engagement and outreach activity
  - Publishing final deliverables as BCP and books
  - Data Science Manifesto – Primarily focused on professional and ethical issues in Data Science, new type of professional
  - Inter-universities initiative “Data Science for UN’s Sustainable Development Goals” to focus in-curricula research (projects) on UN priority goals

- Wider engagement into EOSC activities related to RI Data related skills management and capacity building
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
Data Science Model Curriculum includes

- **Learning Outcomes (LO) definition based on CF-DS**
  - LOs are defined for CF-DS competence groups and for all enumerated competences

- **LOs mapping to Learning Units (LU)**
  - LUs are based on CCS(2012) and universities best practices
  - Data Science university programmes and courses inventory (interactive)
    [http://edison-project.eu/university-programs-list](http://edison-project.eu/university-programs-list)

- **LU/course relevance**: Mandatory Tier 1, Tier 2, Elective, Prerequisite

- **Learning methods and learning models** (in progress)
Example DS-BoK Knowledge Areas definition and mapping to existing BoKs and CCS (2012)

<table>
<thead>
<tr>
<th>Knowledge Area Groups (KAG)</th>
<th>Knowledge Areas (KA)</th>
<th>Suggested Knowledge Units (KU)</th>
<th>Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAG1-DSDA: Data Analytics group (including Machine Learning, statistical methods)</td>
<td>Theory of computation</td>
<td>Machine Learning Theory</td>
<td>CCS2012: Theory of computation Design and analysis of algorithms Data structures design and</td>
</tr>
<tr>
<td>Knowledge Area Groups (KAG)</td>
<td>Knowledge Areas (KA)</td>
<td>Suggested Knowledge Units (KU)</td>
<td>Mapping to CCS2012 (including suggested Data Science extensions) and existing BoKs</td>
</tr>
</tbody>
</table>

- Mapping suggested to CCS2012 and existing BoKs
Example MC-DS Mapping Learning Units to DS-BoK and CCS (2012)

<table>
<thead>
<tr>
<th>KAG/ LU#</th>
<th>Learning Unit (course name)</th>
<th>Type/relevance</th>
<th>Map to DS-BoK, CCS2012 and known BoKs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tier 1 Tier 2 Elective Pre requisite</td>
<td>CCS2012 based academic subjects</td>
</tr>
<tr>
<td>Software requirements and design</td>
<td></td>
<td></td>
<td>DS-BoK and other BoKs</td>
</tr>
<tr>
<td>Information theory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematical analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extensibility point for adding new courses</td>
<td></td>
<td></td>
<td>Mathematical analysis</td>
</tr>
<tr>
<td>Artificial Intelligence</td>
<td></td>
<td></td>
<td>Computing methodologies</td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td></td>
<td></td>
<td>Artificial intelligence</td>
</tr>
</tbody>
</table>

- Mapping suggested to ACM CCS2012, DS-BoK and other related BoKs
Data Science or Data Management Group/Department:

Organisational structure and staffing - EXAMPLE

>> Reporting to CDO/CTO/CEO
• Providing cross-organizational services

• (Managing) Data Science Architect (1)
• Data Scientist (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 data specialists for research institution of 200-300 research staff.

Growing role and demand for Data Stewards and data stewardship
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 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

Domain specific components

- Domain specific data & presentation (visualization)

- Models (and data types)
- Methods
- Processes
- Organisational roles

Data Scientist functions is to translate between two domains

Data Scientist role is to maintain the Data Value Chain (domain specific):
- Data Integration => Organisation/Process/Business Optimisation => Innovation