Profiling Data Science Education and Training

based on
EDISON Data Science Framework (EDSF)

Yuri Demchenko
University of Amsterdam

AACSB Conference on
Data Science at Business Schools
Amsterdam, May 16 / 17, 2017
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
EDSF: How CF-DS was constructed

- Background: Standards and Best Practices
- Jobs market analysis: Demanded Data Science Competences and Skills
Background: Standards and Best Practices

- **e-CFv3.0 - European e-Competence Framework for IT**
  - Structured by 4 Dimensions and organizational processes
    - Competence Areas: Plan – Build – Run – Enable - Manage
    - Competences: total defined 40 competences
    - Proficiency levels: identified 5 levels linked to professional education levels
    - Skills and Knowledge
- **CWA 16458 (2012): European ICT Professional Profiles Family Tree**
  - Defines 23 ICT profiles for common ICT jobs
- **ESCO (European Skills, Competences, Qualifications and Occupations) framework**
  - Standard for European job market since 2016
  - Expected inclusion of the Data Science occupations family – end 2017
- **ACM Classification of Computer Science – CCS (2012)**
- **ACM Computer Science Body of Knowledge (CS-BoK) and ACM and IEEE Computer Science Curricula 2013 (CS2013)**
Background: Standards and Best Practices

- e-CFv3.0 - European e-Competence Framework for IT professionals
  - Structured by 4 Dimensions and organizational processes
    - Competence Areas: Plan, Build, Run, Enable, Manage
    - Key levels: identified 5 levels linked to professional education levels
      - Skills and Knowledge
  - Proficiency levels: identified 5 levels linked to professional education levels

- CWA 16458 (2012): European ICT Professional Profiles Family Tree
  - Defines 23 ICT profiles for common ICT jobs

- ESCO (European Skills, Competences, Qualifications and Occupations) framework
  - Standard for European job market since 2016
  - Expected inclusion of the Data Science occupations family – end 2017

- ACM Classification of Computer Science – CCS (2012)
  - New Joint Initiative ACM, IEEE, ASA, AAAS, AIS, ACH
  - To develop Data Science curriculum

Currently work on e-CF4 is moved to CEN TC 428
To be extended with Data Science competences
Jobs market analysis: Demanded Data Science Competences and Skills

• Initial Analysis (period Aug – Sept 2015) -> Continuous monitoring (in development)
  – 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 & forums

• Analysis methods
  – Data analytics methods: classification, clustering, feature extraction
  – Research methods: Data collection - Hypothesis – Artefact - Evaluation
  – Expert evaluation by EDISON Liaison Groups (ELG), multiple workshops
**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

Amsterdam, 17 May 2017
Data Science Competences include 5 groups:
- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Scientific Methods or Business Process Management

Business Process Operations/Stages:
- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design

Scientific Methods:
- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis
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**
  – Critical thinking, personal, inter-personal communication, team work, professional network
## Identified Data Science Competence Groups

<table>
<thead>
<tr>
<th>Data Science Analytics (DSDA)</th>
<th>Data Management (DSDM)</th>
<th>Data Science Engineering (DSENG)</th>
<th>Research/Scientific Methods (DSRM)</th>
<th>Data Science Domain Knowledge, e.g. Business Processes (DSDK/DSBPM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations</td>
<td>Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.</td>
<td>Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management</td>
<td>Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals</td>
<td>Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations</td>
</tr>
<tr>
<td>DSDA01 Use predictive analytics to analyse big data and discover new relations</td>
<td>DSDM01 Develop and implement data strategy, in particular, Data Management Plan (DMP)</td>
<td>DSENG01 Use engineering principles to design, prototype data analytics applications, or develop instruments, systems</td>
<td>DSRM01 Create new understandings and capabilities by using scientific/research methods or similar domain related development methods</td>
<td>DSBPM01 Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework</td>
</tr>
<tr>
<td>DSDA02 Use statistical techniques to deliver insights</td>
<td>DSDM02 Develop data models including metadata</td>
<td>DSENG02 Develop and apply computational solutions</td>
<td>DSRM02 Direct systematic study toward a fuller knowledge or understanding of the observable facts</td>
<td>DSBPM02 Participate strategically and tactically in financial decisions</td>
</tr>
<tr>
<td>DSDA03 Develop specialized applications</td>
<td>DSDM03 Collect integrate data</td>
<td>DSENG03 Develops specialized tools</td>
<td>DSRM03 Undertakes creative work</td>
<td>DSBPM03 Provides support services to other</td>
</tr>
<tr>
<td>DSDA04 Analyze complex data</td>
<td>DSDM04 Maintain repository</td>
<td>DSENG04 Design, build, operate</td>
<td>DSRM04 Translate strategies into actions</td>
<td>DSBPM04 Analyse data for marketing</td>
</tr>
<tr>
<td>DSDA05 Use different analytics</td>
<td>DSDM05 Visualise complex data</td>
<td>DSENG05 Secure and reliable data</td>
<td>DSRM05 Contribute to organizational goals</td>
<td>DSBPM05 Analyse optimise customer relation</td>
</tr>
</tbody>
</table>
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.
DSP Profiles mapping to ESCO Taxonomy
High Level Groups and CF-DS Competences

- DSP Profiles mapping to corresponding CF-DS Competence Groups
  - Relevance level from 5 – maximum to 1 – minimum
Education and Training – Part of EDSF

• 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 to Knowledge Units (KU) and Learning Units (LU)

- EDSF allow for customized educational courses and training modules design
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</td>
<td>Theory of computation</td>
<td>Design and Analysis of Algorithms</td>
<td>CCS2012: Theory of computation Design and analysis of algorithms Data structures design and</td>
</tr>
<tr>
<td>Data Analytics group</td>
<td></td>
<td>Machine Learning Theory</td>
<td>Machine Learning Theory</td>
</tr>
<tr>
<td>(including Machine Learning, statistical methods)</td>
<td></td>
<td></td>
<td>Machine Learning Theory</td>
</tr>
<tr>
<td>KAG2: DSENG</td>
<td>Computer systems organisation for Big Data</td>
<td>Parallel and Distributed Computer Architecture</td>
<td>CCS2012: Computer systems organization Architectures Parallel architectures</td>
</tr>
<tr>
<td>Data Science Engineering group including Software and Infrastructures Engineering</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- Mapping suggested to ACM CCS2012, DS-BoK and other related BoKs
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
  - Followed by Ministerial meeting on 14-15 May 2017 in Hanoi, Vietnam

- **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

- **e-IRG workshop on Sustainability** on 8-9 June 2017, Malta
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/](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 yrs
  - 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
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)
- **Community Portal** [http://datasciencepro.eu/](http://datasciencepro.eu/)

- 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/

• EDISON Data Science Framework Release 1 (EDSF)
  http://edison-project.eu/edison-data-science-framework-edsf
    – Data Science Competence Framework
      http://edison-project.eu/data-science-competence-framework-cf-ds
    – Data Science Body of Knowledge
      http://edison-project.eu/data-science-body-knowledge-ds-bok
    – Data Science Model Curriculum
      http://edison-project.eu/data-science-model-curriculum-mc-ds
    – Data Science Professional Profiles
      http://edison-project.eu/data-science-professional-profiles-definition-dsp

• Survey Data Science Competences: Invitation to participate
  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)