Defining a new profession of the Data Scientist to address critical skills gap in European research and industry

Yuri Demchenko, EDISON
University of Amsterdam

Cloud and DevOps World 2016

21-22 June 2016, Olympia, London, UK
Outline

• Background and motivation
  – Demand for Data Science and data related professions
  – European initiatives related to Digital Single Market (DSM) and demand to data related competences and skills

• EDISON Data Science Framework
  – From Data Science Competences to Body of Knowledge and Model Curriculum

• Data Science Competence Framework: Essential competences and skills

• Taxonomy: Data Science professions family

• Data Science Body of Knowledge (DS-BoK)
  • Knowledge areas and academic disciplines

• Further steps
Demand for Data Science and data related professions

- McKinsey Global Institute on Big Data Jobs (2011)
  - Estimated gap of 140,000 - 190,000 data analytics skills by 2018

- UK Big Data skills report 2014
  - 6400 UK organisations with 100+ staff will have implemented Big Data Analytics by 2020
  - Increase of Big Data jobs from 21,400 (2013) to 56,000 (2017)

  - 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%

- HLEG report on European Open Science Cloud (2016) identified need for data experts and data stewards
  - Recommendation: Allocate 5% from grant funding for Data management and preservation
  - Estimation: More than 80,000 data stewards (1 per every 20 scientists)
  - Core data experts need to be trained and their career perspective improved
Recent European Commission Initiatives


- The need for new multidisciplinary and digital skills is exploding, including such as (Data Scientist) combining data analytics and business or engineering skills.
  - Expected rapidly growing demand will lead to more than 800,000 unfilled vacancies by 2020.
- The forthcoming New Skills Agenda for Europe (exp May 2016) to address the need for digital and complementary skills, ensure young talents flow into data driven research and industry.


- **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.
- Raise awareness and change incentive structures for academics, industry and public services to share their data, and improve data management training, literacy and data stewardship skills.
- Address growing demand and shortage of data-related skills and lack of recognition of their value (in all sectors).
• EDISON Framework components
  • CF-DS – Data Science Competence Framework
  • DS-BoK – Data Science Body of Knowledge
  • MC-DS – Data Science Model Curriculum
  • Data Science Taxonomies and Scientific Disciplines Classification
    – Linked to e-CFv3.0, ACM CCS (2012) and ESCO
  • EOEE - EDISON Online Education Environment
Background Frameworks and Standards

• 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
  – Intended inclusion of the Data Science occupations family – end of 2016

• ACM Classification of Computer Science – CCS (2012)
  – ACM Computer Science Body of Knowledge (CS-BoK) and ACM and IEEE Computer Science Curricula 2013 (CS2013)
Data Science Competence Framework is a foundation for the Data Science Profession definition

- How it was made
- 5 main Data Science competences groups
- Skills, tools and languages
- Practical use of the CF-DS (examples)
  - Suggested e-CFv3.0 extensions for Data Science
  - Data Science occupations family
Demanded Data Science Competences and Skills: Jobs market analysis

• Initial Analysis (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 & forums

• 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 without specified skills profile
  – Explicit Data Scientist jobs specify wide variety of expected functions/responsibilities and required skills and knowledge
**Data Scientist definition by NIST**

Definitions by NIST Big Data WG (NIST SP1500 - 2015)

- **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.

- **Data science** is the empirical synthesis of actionable knowledge and technologies required to handle data from raw data through the complete data lifecycle process.

[ref] Legacy: NIST BDWG definition of Data Science
Identified Data Science Competence Groups

• Commonly accepted 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/or 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 programmes
  – 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 and understandable way of communication and information/data presentation
  – Role of Data Scientist: Provide such literacy advice and guiding to organisation
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 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
<table>
<thead>
<tr>
<th>Identified Data Science Competence Groups (Updated)</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></td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
• **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
### Data Science Skill Groups related to Competences

<table>
<thead>
<tr>
<th>Data Analytics and Machine Learning</th>
<th>Data Management/ Curation</th>
<th>Data Science Engineering (hardware and software)</th>
<th>Scientific/ Research Methods</th>
<th>Personal/Inter-personal communication, team work</th>
<th>Application/subject domain (research or business), examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Artificial intelligence, machine learning</td>
<td>Manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources</td>
<td>Design efficient algorithms for accessing and analyzing large amounts of data</td>
<td>Analytical, independent, critical, curious and focused on results</td>
<td>Communication skills</td>
<td>Recommender or Ranking system</td>
</tr>
<tr>
<td>2 Machine Learning and Statistical Modelling</td>
<td>Data models and datatypes</td>
<td>Big Data solutions and advanced data mining tools</td>
<td>Confident with large data sets and ability to identify appropriate tools and algorithms</td>
<td>Inter-personal intra-team and external communication</td>
<td>Data Analytics for commercial purposes</td>
</tr>
<tr>
<td>3 Machine learning solutions and pattern recognition techniques</td>
<td>Experience of working with large data sets</td>
<td>Multi-core/distributed software, preferably in a Linux environment</td>
<td>Flexible analytic approach to achieve results at varying levels of precision</td>
<td>Network of contacts in Big Data community</td>
<td>Data sources and techniques for business insight and customer focus</td>
</tr>
<tr>
<td>4 Supervised and unsupervised learning</td>
<td>(non)relational and (un)-structured data</td>
<td>Databases, database systems, SQL and NoSQL</td>
<td>Interest in data science, exceptional analytical skills</td>
<td>Mechanism Design and/or Latent Dirichlet Allocation</td>
<td></td>
</tr>
<tr>
<td>5 Data mining</td>
<td>Cloud based data storage and data management</td>
<td>Statistical analysis languages and tooling</td>
<td></td>
<td>Game Theory</td>
<td></td>
</tr>
<tr>
<td>6 Markov Models, Conditional Random Fields</td>
<td>Data management planning</td>
<td>Cloud powered applications design</td>
<td></td>
<td>Copyright and IPR</td>
<td></td>
</tr>
<tr>
<td>7 Logistic Regression, Support Vector Machines</td>
<td>Metadata annotation and management</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Predictive analysis and statistics (including Kaggle platform)</td>
<td>Data citation, metadata, PID (*)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Artificial Neural Networks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Mathematics foundation:** Knowledge of mathematics, calculus, probability theory and statistics
## Identified Big Data Tools and Programming Languages

<table>
<thead>
<tr>
<th>Big Data Analytics platforms</th>
<th>Math&amp; Stats tools</th>
<th>Databases</th>
<th>Data/ applications visualization</th>
<th>Data Management and Curation platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Big Data Analytics platforms</td>
<td>Advanced analytics tools (R, SPSS, Matlab, etc)</td>
<td>SQL and relational databases</td>
<td>Data visualization Libraries (D3.js, FusionCharts, Chart.js, other)</td>
<td>Data modelling and related technologies (ETL, OLAP, OLTP, etc)</td>
</tr>
<tr>
<td>2 Big Data tools (Hadoop, Spark, etc)</td>
<td>Data Mining tools: RapidMiner, others</td>
<td>NoSQL Databases</td>
<td>Visualisation software (D3, Processing, Tableau, Gephi, etc)</td>
<td>Data warehouses platform and related tools</td>
</tr>
<tr>
<td>3 Distributed computing tools a plus (Spark, MapReduce, Hadoop, Hive, etc.)</td>
<td>Mathlab</td>
<td>NoSQL, Mongo, Redis</td>
<td>Online visualization tools (Datawrapper, Google Charts, Flare, etc)</td>
<td>Data curation platform, metadata management (ETL, Curator’s Workbench, DataUp, MIXED, etc)</td>
</tr>
<tr>
<td>4 Real time and streaming analytics systems (like Flume, Kafka, Storm)</td>
<td>Python</td>
<td>NoSQL, Teradata</td>
<td>Backup and storage management (iRODS, XArch, Nesstar, others)</td>
<td></td>
</tr>
<tr>
<td>5 Hadoop Ecosystem/platform</td>
<td>R, Tableau R</td>
<td>Excel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Spotfire</td>
<td>SAS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Azure Data Analytics platforms (HDInsight, APS and PDW, etc)</td>
<td>Scripting language, e.g. Octave</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Amazon Data Analytics platform (Kinesis, EMR, etc)</td>
<td>Statistical tools and data mining techniques</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Other cloud based Data Analytics platforms, e.g. HortonWorks, Vertica LexisNexis HPCC System</td>
<td>Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Highlighted:**
Cloud based and online data analytics and data management platforms
Suggested 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 (including CV matching)
  - For customizable training and career development
- 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
Suggested e-CF Competences for Data Science: Next eCF Workshop meeting – Presented 14 April 2016

A. PLAN and Design (9 basic competences)
   • A.10* Organisational workflow/processes model definition/formalisation
   • A.11* Data models and data structures

B. BUILD: Develop and Deploy/Implement (6 basic competences)
   • 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 (4 basic competences)
   • C.5* User/Usage data/statistics analysis
   • C.6* Service delivery/quality data monitoring

D. ENABLE: Use/Utilise (12 basic competences)
   • D.10. 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 (9 basic competences)
   • 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)

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

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

- Data Integration $\Rightarrow$ Organisation/Process/Business Optimisation $\Rightarrow$ **Innovation**
Data Scientist role is to maintain the Data Value Chain (domain specific):
- Data Integration => Organisation/Process/Business Optimisation => Innovation
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 and clerical (data handling/management):** Data Stewards, Digital Data Curator, Digital Librarians, Data Archivists

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

Icons used: Credit to [ref] https://www.datacamp.com/community/tutorials/data-science-industry-infographic
### Data Science Occupations: Extension for the ESCO (2016) taxonomy (1)

<table>
<thead>
<tr>
<th>Professionals</th>
<th>Data Science Professionals</th>
<th>Data Science professionals not elsewhere classified</th>
<th>DSP#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science and engineering professionals</td>
<td>Data Science Professionals</td>
<td>Data Science professionals not elsewhere classified</td>
<td>DSP04 Data Scientist</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DSP05 Data Science Researcher</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DSP08 (Big) Data Analyst</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DSP07 Data Science (Application) Programmer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DSP09 Business Analyst</td>
</tr>
<tr>
<td></td>
<td>Database and network professionals</td>
<td>Large scale (cloud) data storage designers and administrators</td>
<td>DSP14 Large scale (cloud) database designer*</td>
</tr>
<tr>
<td></td>
<td>Database designers and administrators</td>
<td>Database designers and administrators</td>
<td>DSP15 Large scale (cloud) database administrator*</td>
</tr>
<tr>
<td></td>
<td>Database and network professionals not elsewhere classified</td>
<td>Data handling professionals not elsewhere classified</td>
<td>DSP12 Digital Librarian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DSP13 Data Archivist</td>
</tr>
<tr>
<td>Information and communications technology professionals</td>
<td>Data Science technology professionals</td>
<td>Data handling professionals not elsewhere classified</td>
<td>DSP10 Data Steward</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DSP11 Data curator</td>
</tr>
</tbody>
</table>

19 DSP# Enumerated Data Science profiles defined by EDISON Framework
## Data Science Occupations: Extension for the ESCO taxonomy (2)

<table>
<thead>
<tr>
<th>Technicians and associate professionals</th>
<th>Data Science Technology Professionals</th>
<th>Data Infrastructure engineers and technicians</th>
<th>DSP17 Big Data facilities Operators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science and engineering associate professionals</td>
<td></td>
<td></td>
<td>DSP18 Large scale (cloud) data storage operators</td>
</tr>
<tr>
<td>Database and network professionals not elsewhere classified</td>
<td></td>
<td></td>
<td>DSP19 Scientific database operator*)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Managers</th>
<th>Data Science/Big Data Infrastructure Managers</th>
<th>DSP01/DSP02 Data Science/Big Data Infrastructure Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production and specialised services managers</td>
<td></td>
<td>DSP03 RI Manager</td>
</tr>
<tr>
<td>Research Infrastructure Managers</td>
<td></td>
<td>DSP03 RI Data storage facilities manager</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clerical support workers</th>
<th>Data handling support workers (alternative)</th>
<th>Data and information entry and access</th>
<th>Digital Archivists and Librarians</th>
<th>DSP12* Digital Librarian</th>
</tr>
</thead>
<tbody>
<tr>
<td>General and keyboard clerks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data handling support workers (alternative)</td>
<td>Data and information entry and access</td>
<td>Digital Archivists and Librarians</td>
<td>DSP12* Digital Librarian</td>
<td></td>
</tr>
<tr>
<td>DSP13* Data Archivist</td>
<td>DSP10* Data Steward</td>
<td>DSP11* Data curator</td>
<td>DSP12* Digital Librarian</td>
<td></td>
</tr>
</tbody>
</table>
Data Science or Data Management Group/Department

- Group Manager
- Data Science Architect
- Data Analyst
- Data Science Application programmer
- Data Infrastructure/facilities administrator/operator: storage, cloud, computation
- Data stewards

>> Reporting to CDO/CTO/CEO
   - Providing cross-organizational services
   - Maintaining Data Value Chain
Data Science or Data Management Group/Department: Organisational structure and staffing - EXAMPLE

Data Science or Data Management Group/Department

>> Reporting to CDO/CTO/CEO
  • Providing cross-organizational services
  • Maintaining Data Value Chain

• (Managing) Data Science Architect (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 specialists for research institution of 200-300 staff.
Education and Training

• Foundation and methodological base
  – Data Science Body of Knowledge (DS-BoK)
    • Taxonomy and classification of Data Science related scientific subjects
  – Instructional methodologies and teaching models

• Platforms and environment
  – Virtual labs, datasets, developments platforms
  – Access control and accounts/identity management
  – Online education environment and courses management

• Services
  – Individual benchmarking and profiling tools (competence assessment)
  – Knowledge evaluation tools
  – Certifications and training
  – Education and training marketplace: Courses catalog and repository
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
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)
Research Data Management Model Curriculum – Part of the EDISON Data Literacy Training

A. Use cases for data management and stewardship
   • Preserving the Scientific Record

B. Data Management elements (organisational and individual)
   • Goals and motivation for managing your data
   • Data formats
   • Creating documentation and metadata, metadata for discovery
   • Using data portals and metadata registries
   • Tracking Data Usage
   • Handling sensitive data
   • Backing up your data
   • Data Management Plan (DMP) - to be a part of hands on session

C. Responsible Data Use Section (Citation, Copyright, Data Restrictions)

D. Open Science and Open Data (Definition, Standards, Open Data use and reuse, open government data)
   • Research data and open access
   • Repository and self-archiving services
   • ORCID identifier for data
   • Stakeholders and roles: engineer, librarian, researcher
   • Open Data services: ORCID.org, Altmetric Doughnut, Zenodo

E. Hands on:
   a) Data Management Plan design
   b) Metadata and tools
   c) Selection of licenses for open data and contents (e.g. Creative Common and Open Database)
Further Steps

- Define a taxonomy and classification for DS competences and skills as a basis for more formal CF-DS definition
- 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
- Formally provide suggestions for e-CF3.0 extensions for Data Science to CEN/PC 428
  - Involve national e-CF bodies and adopters where available
- Provide feedback and contribution to ESCO with the definition of the Data Science professions family
- Suggest ACM CCS2012 Classification extensions and officially contact ACM
- Involve academic and industry experts and professional organisations in the definition of DS-BoK following from CF-DS
  - Link to taxonomy of academic and educational and training courses
- Invite and analyse community contribution main EDISON deliverables
  - CF-DS, DS-BoK, DS Professional Profiles documents are on public comments and available from the EDISON website
Discussion

- Questions
- Observations
- Suggestions

- Survey Data Science Competences [1]: Invitation to participate
  https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession

- Community discussion documents: Request for comments
  - 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 Professional Profiles
    http://edison-project.eu/data-science-professional-profiles
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 Approach: CF-DS and e-CFv3.0

- 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)*

- 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
### European e-Competence Framework 3.0 overview

<table>
<thead>
<tr>
<th>Dimension 1 5 e-CF areas (A – E)</th>
<th>Dimension 2 40 e-Competences identified</th>
<th>Dimension 3 e-Competence proficiency levels e-1 to e-5, related to IQP levels 3 – 8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. PLAN</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.1. IS and Business Strategy Alignment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.2. Service Level Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.3. Business Plan Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.4. Product/Service Planning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.5. Architecture Design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.6. Application Design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.7. Technology Trend Monitoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.8. Sustainable Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A.9. Innovating</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B. BUILD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.1. Application Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.2. Component Integration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.3. Testing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.4. Solution Deployment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.5. Documentation Production</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.6. Systems Engineering</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C. RUN</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.1. User Support</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.2. Change Support</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.3. Service Delivery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.4. Problem Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D. ENABLE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.1. Information Security Strategy Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.2. ICT Quality Strategy Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.3. Education and Training Provision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.4. Purchasing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.5. Sales Proposal Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.6. Channel Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.7. Sales Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.8. Contract Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.9. Personnel Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.10. Information and Knowledge Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.11. Needs Identification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.12. Digital Marketing</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>E. MANAGE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E.1. Forecast Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E.2. Project and Portfolio Management</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Consider for Data Science:
- 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)