EDISON Framework: 
Data Science Competence Framework (CF-DS) 
and 
Data Science Body of Knowledge (DS-BoK)
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: Data Science occupations Family (proposed ESCO extension)
- Data Science Body of Knowledge (DS-BoK)
  - Taxonomy: Knowledge area, academic disciplines
- Further steps - Survey and questionnaires
EOEE - EDISON Online Education Environment

• 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

Sustainability:
• Roadmap
• Community Portal
• Certification
• Professional community/association
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)
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

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
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

<table>
<thead>
<tr>
<th>Data Analytics (DA)</th>
<th>Data Management/ Curation (DM)</th>
<th>DS Engineering (DSE)</th>
<th>Search Methods (DSRM) scientific/Re</th>
<th>DS Domain Knowledge (including Business Apps)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>Use appropriate statistical techniques on available data to deliver insights</td>
<td>Develop and implement data strategy</td>
<td>Use engineering principles to research, design, or develop structures, instruments, machines, experiments, processes, systems, theories, or technologies</td>
<td>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</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>Use predictive analytics to analyse big data and discover new relations</td>
<td>Develop data models including metadata</td>
<td>Develops specialized data analysis tools to support executive decision making</td>
<td>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</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>Research and analyze complex data sets, combine different sources and types of data to improve analysis.</td>
<td>Integrate different data source and provide for further analysis</td>
<td>Design, build, operate relational non-relational databases</td>
<td>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</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>Develop specialized analytics to enable agile decision making</td>
<td>Develop and maintain a historical data repository of analysis</td>
<td>Develop and apply computational solutions to domain related problems using wide range of data analytics platforms</td>
<td>Apply ingenuity to complex problems, develop innovative ideas</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>Collect and manage different source of data</td>
<td>Collect and manage different source of data</td>
<td>Develop solutions for secure and reliable data access</td>
<td>Ability to translate strategies into action plans and follow through to completion.</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>Visualise complex and variable data.</td>
<td>Develop algorithms to analyse multiple source of data</td>
<td>Analyse multiple data sources for marketing purposes</td>
<td>Influences the development of organizational objectives</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>Prototype new data analytics applications</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**IG-ETRD @ RDA7**
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
  – Application/subject domain related (research or business)
  – Mathematics and Statistics

• **Group 2: Big Data (Data Science) tools and platforms**
  – Big Data Analytics platforms
  – Math & Stats apps & 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 (also called social intelligence or soft skills)
## Identified Data Science Skill Groups

<table>
<thead>
<tr>
<th>Identified Data Science Skill Groups</th>
<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)</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 for data improvement</td>
<td>Design efficient algorithms for accessing and analyzing large amounts of data</td>
<td>Interest in data science</td>
<td>Communication skills</td>
<td>Recommender or Ranking system</td>
<td></td>
</tr>
<tr>
<td>2 Machine Learning and Statistical Modelling</td>
<td>Big Data solutions and advanced data mining tools</td>
<td>Analytical, independent, critical, curious and focused on results</td>
<td>Inter-personal intra-team and external communication</td>
<td>Data Analytics for commercial purposes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Machine learning solutions and pattern recognition techniques</td>
<td>Data models and datatypes</td>
<td>Multi-core/distributed software, preferably in a Linux environment</td>
<td>Confident with large data sets and ability to identify appropriate tools and algorithms</td>
<td>Network of contacts in Big Data community</td>
<td>Data sources and techniques for business insight and customer focus</td>
<td></td>
</tr>
<tr>
<td>4 Supervised and unsupervised learning</td>
<td>Handling vast amounts of data</td>
<td>Databases, database systems, SQL and NoSQL</td>
<td>Flexible analytic approach to achieve results at varying levels of precision</td>
<td>Mechanism Design and/or Latent Dirichlet Allocation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Data mining</td>
<td>Experience of working with large data sets</td>
<td>Statistical analysis languages and tooling</td>
<td>Exceptional analytical skills</td>
<td>Game Theory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Markov Models, Conditional Random Fields</td>
<td>(non)relational and (un)-structured data</td>
<td>Cloud powered applications design</td>
<td></td>
<td>Copyright and IPR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Logistic Regression, Support Vector Machines</td>
<td>Cloud based data storage and data management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Predictive analysis and statistics (including Kaggle platform)</td>
<td>Data management planning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 (Artificial) Neural Networks</td>
<td>Metadata annotation and management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Statistics</td>
<td>Data citation, metadata, PID (*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

*IG-ETRD @ RDA7 Data Science Competences and BoK*
## 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></td>
<td>Backup and storage management (iRODS, XArch, Nesstar, others)</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 (HortonWorks, Vertica LexisNexis HPCC System, etc)</td>
<td>Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Big Data Analytics platforms
- R, SPSS, Matlab, etc
- D3.js, FusionCharts, Chart.js, other
- ETL, OLAP, OLTP, etc

### Math & Stats tools
- Advanced analytics tools
- Data Mining tools: RapidMiner, others
- Mathlab
- Scripting language, e.g. Octave
- Statistical tools and data mining techniques
- Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)

### Databases
- SQL and relational databases
- NoSQL Databases
- NoSQL, Mongo, Redis
- NoSQL, Teradata

### Data/ applications visualization
- Data visualization Libraries (D3.js, FusionCharts, Chart.js, other)
- Visualisation software (D3, Processing, Tableau, Gephi, etc)
- Online visualization tools (Datawrapper, Google Charts, Flare, etc)

### Data Management and Curation platform
- Data modelling and related technologies (ETL, OLAP, OLTP, etc)
- Data warehouses platform and related tools
- Data curation platform, metadata management (ETL, Curator's Workbench, DataUp, MIXED, etc)
- Backup and storage management (iRODS, XArch, Nesstar, others)
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

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)
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

Data Science domain components

- Abstract data driven math&compute models
- Data Analytics methods
- Data and Applications Lifecycle Management

Domain specific components

- Domain specific data & presentation
- Models (and data types)
- Methods
- Processes

Organisational roles

Data Scientist functions is to translate between two domains

Cross-organisational assistive role
Possible Data Scientist profiles/roles as extension to CWA16458 (2012)

• **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 programming
  – Data Science/Big Data Infrastructure development/operation

• **Data Science Researcher**
  – Data Science research methods
  – Data models and structures

• **Data Scientist in subject/research domain**

• **Research e-Infrastructure brings its own specifics to required competences and skills definition**
## Proposed Placing of Data Science Occupations Family in ESCO taxonomy (1)

<table>
<thead>
<tr>
<th>Professionals</th>
<th>Data Science Professionals</th>
<th>Data Science professionals not elsewhere classified</th>
<th>Data Scientist</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>Data Science Researcher</td>
</tr>
<tr>
<td>(Big) Data Analyst</td>
<td></td>
<td></td>
<td>(Big) Data Analyst</td>
</tr>
<tr>
<td>Data Science (Application) Programmer</td>
<td></td>
<td></td>
<td>Data Science (Application) Programmer</td>
</tr>
<tr>
<td>Business Analyst</td>
<td></td>
<td></td>
<td>Business Analyst</td>
</tr>
<tr>
<td>Database and network professionals</td>
<td>Large scale (cloud) data storage designers and administrators</td>
<td>Large scale (cloud) database designer*</td>
<td>Large scale (cloud) database designer*</td>
</tr>
<tr>
<td>Database designers and administrators</td>
<td></td>
<td></td>
<td>Database designers and administrators</td>
</tr>
<tr>
<td>Database and network professionals not elsewhere classified</td>
<td>Scientific database administrator*</td>
<td></td>
<td>Scientific database administrator*</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>Digital Librarian</td>
</tr>
<tr>
<td>Digital Librarian</td>
<td></td>
<td></td>
<td>Digital Librarian</td>
</tr>
<tr>
<td>Data Archivist</td>
<td></td>
<td></td>
<td>Data Archivist</td>
</tr>
<tr>
<td>Data Steward</td>
<td></td>
<td></td>
<td>Data Steward</td>
</tr>
<tr>
<td>Data curator</td>
<td></td>
<td></td>
<td>Data curator</td>
</tr>
</tbody>
</table>
## Proposed Placing of Data Science Occupations Family in ESCO taxonomy (2)

### Technicians and associate professionals

| Science and engineering associate professionals | Data Science Technology Professionals | Data Infrastructure engineers and technicians | Big Data facilities Operators
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Science and engineering associate professionals</td>
<td>Data Science Technology Professionals</td>
<td>Data Infrastructure engineers and technicians</td>
<td>Big Data facilities Operators</td>
</tr>
</tbody>
</table>

### Managers

<table>
<thead>
<tr>
<th>Production and specialised services managers</th>
<th>Data Science/Big Data Infrastructure Managers</th>
<th>Data Science/Big Data Infrastructure Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production and specialised services managers</td>
<td>Data Science/Big Data Infrastructure Managers</td>
<td>Data Science/Big Data Infrastructure Manager</td>
</tr>
</tbody>
</table>

### Clerical support workers

<table>
<thead>
<tr>
<th>General and keyboard clerks</th>
<th>Data handling support workers (alternative)</th>
<th>Data and information entry and access</th>
<th>Digital Archivists and Librarians</th>
<th>Digital Librarian</th>
</tr>
</thead>
<tbody>
<tr>
<td>General and keyboard clerks</td>
<td>Data handling support workers (alternative)</td>
<td>Data and information entry and access</td>
<td>Digital Archivists and Librarians</td>
<td>Digital Librarian</td>
</tr>
<tr>
<td>General and keyboard clerks</td>
<td>Data handling support workers (alternative)</td>
<td>Data and information entry and access</td>
<td>Digital Archivists and Librarians</td>
<td>Digital Librarian</td>
</tr>
</tbody>
</table>

- Database and network professionals not elsewhere classified
- Large scale (cloud) data storage operators
- Scientific database operator*
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
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 European Open Data initiatives, European Open Data Cloud, and RDA (Research Data Alliance)

(12) PID, metadata, data registries
(13) Data Management Plan
(14) Open Science, Open Data, Open Access, ORCID
(15) Responsible data use
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)
• 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
  - Start related Social Network groups to promote already obtained results and obtain feedback and community contribution
## 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? |
|-----------------------------------------------|-----------------------------------------------|
| **Factually and theoretically knowledge**    | **Comprehensive factually and theoretically knowledge** | **Advanced knowledge of a field, critical understanding of theories and principles** | **Highly specialized knowledge** | **Critical awareness, interfere between different fields** | **Knowledge at the most advanced frontier of a field** |
| Use appropriate statistics to provide insight on data |                                |                                    |                                |                                |                                |
| Use appropriate techniques for analyzing 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 and sources of data, to improve analysis |                                |                                    |                                |                                |                                |
| Develop specialized analytics to enable agile decision making |                                |                                    |                                |                                |                                |