EDISON Project Overview:

Building the Data Science Profession for Research and Industry

Yuri Demchenko, University of Amsterdam

2nd EDISON Data Science Champions Conference
15-16 March 2017
Universidad Carlos III de Madrid, Madrid
Outline

- Background and motivation
  - European initiatives related to Digital Single Market and Digital Skills Agenda
  - EDISON building network to promote establishing Data Science profession
- EDISON Data Science Framework (EDSF)
  - From Data Science Competences to Body of Knowledge and Model Curriculum
- Data Science Competence Framework: Essential competences and skills
- Education and training focus
  - Data Science Body of Knowledge (DS-BoK)
  - Data Science Model Curriculum (MC-DS)
  - Example Research Data Management Literacy curriculum
- Further developments to formalise EDSF and Data Science profession establishment
EDISON Objectives, Impact and Actions

**Impact**

Increase the number of Data Scientists and Market for establishing Data Science Profession

**Objectives and Actions**

- **Create a Data Science profession**
  - Data Science professional profiles
  - Interact with demand and supply sides
  - Career path building and skills transferability

- **Services to education and training**
  - Define Model Curriculum and design tools
  - Support for accreditation and certification
  - Collaborating and sharing expertise and materials

- **Engage stakeholder communities**
  - Sustain platforms of communities of practice
  - Create community of “champion” universities
  - Interact with Expert Liaison Groups

**Data Science Competence Framework and Body of Knowledge**
Visionaries and Drivers:
Seminal works, High level reports, Activities

The Fourth Paradigm: Data-Intensive Scientific Discovery.

Riding the wave: How Europe can gain from the rising tide of scientific data.

The Data Harvest: How sharing research data can yield knowledge, jobs and growth.
An RDA Europe Report. December 2014

HLEG report on European Open Science Cloud
(October 2016)

Emergence of Cognitive Technologies
(IBM Watson and others)
Recent European Commission Initiatives 2016

- 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

- **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
- Address growing demand and shortage of data-related skills

- Addresses the need for digital and complementary skills, ensure young talents flow into data driven research and industry
- Launch **Digital Skills and Jobs Coalition** (1st December 2016, Brussels) to develop comprehensive national digital skills strategies by mid-2017


• Definition of the Data Steward as a distinctive role and profession
  – Core Data Experts need to be trained and their career perspective improved

• Estimation: More than 80,000 data stewards to serve 1.7 mln scientists in Europe (1 per every 20 scientists)
  – Based on 5% grant funding for Data management and preservation

• Clash of cultures between domain specialists and e-Infrastructure specialists (i.e. IT/Computer Science)
OECD

• Demand for new type of “dynamic self-re-skilling workforce”
• Continuous learning and professional development to become a shared responsibility of workers and organisations

[ref] SKILLS FOR A DIGITAL WORLD, OECD, 25-May-2016

UN

• Data Revolution Report "A WORLD THAT COUNTS" Presented to Secretary-General (2014)
  http://www.undatarevolution.org/report/
• Data Literacy is defined as key for digital revolution
• Data literacy = critically analyse data collected and data visualised
Approach

- Task is not for one project – **Need collaboration**
- Task is not for science only in isolation from industry
- Needs strong **conceptual approach**
  - Use science to solve the problems of science
- **Standardisation** is an important factor of sustainability and development
EDISON Network and Engagement Activity (1)

- Cooperative relations and exchange of developments with RI projects
  - ELIXIR, CORBEL/RItrain, EUDAT, ENVRI
  - FOSTER2, EOSCpilot
- Cooperation with Big Data and Data Science projects
  - EDSA, BDVA
- Active contribution to the Research Data Alliance (RDA) activities
  - RDA IG on Education and Training on Handling Research Data (IG-ETHRD)
  - BoFs and proposed WG on Certification and accreditation,
  - Proposed WG on Text Data Mining
  - Proposed WG on Research Data Management Curriculum
  - BoF on Data Champions
Workshops to promote a common approach towards addressing growing demand for Data Science and critical data competences and skills as required by European Research Infrastructures (RI), future European Open Science Cloud (EOSC) and generally European Digital Single Market (DSM).

- **Joint EDISON and EC workshop** “Data Infrastructure Competences and Skills Framework: a European and Global Challenge” (Brussels, 9th February, 2016)
- Joint IEEE, STC CC and RDA Workshop on Curricula and Teaching Methods (DTW2015 and DTW2016 collocated with IEEE CloudCom)

EDISON initiated a set of **national action meetings** to address Data Science and digital skills by bringing together key stakeholders from universities, employer associations, and government.

- A first workshop jointly organised by the EDISON project and Dutch Ministry of Education, Culture and Science in June 2016 (during Netherlands Presidency in EU)
EDISON Network and Engagement Activity (3)

- European and international standardisation bodies and professional organisations
  - CEN TC426 Committee (former e-Competence Framework e-CFv3.0 workshop)
  - ESCO (European Skills, Competence, Occupations)
  - CEPIS and association ICT Professionalism Europe (co-signed 21 Nov 2016, Amsterdam)

- EDISON Booth at the Launch event of the Digital Skills and Jobs Coalition: Boosting Europe's Digital skills, 1 December 2016, Brussels
  - Part of actions toward European Digital Single Market (DSM) and

- Contribution to International standardization bodies, professional organisations and initiatives
  - Business Higher Education Forum (BHEF) in USA
  - DARE project for APEC (Asia Pacific Economic Cooperation) to develop a Data Analytics checklist for APEC countries
  - Data Science Curriculum Meeting of Professional and Academic Societies in USA (4 March 2017, Alexandria) including ACM
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
- Job market study, existing practices in academic, research and industry.
- Compliance with related standards
- 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 expected scientific and business value to science or industry

- **Other definitions to admit such features as**
  - Ability to solve variety of business problems
  - Optimize performance and suggest new services for the organisation
  - Develop a special mindset and be statistically minded, *understand raw data* and “*appreciate data as a first class product*”

- **Data science** is the empirical synthesis of actionable knowledge and technologies required to handle data from raw data through the complete data lifecycle process.
- **Big Data** is the technology to build system and infrastructures to process large volume of structurally complex data in a time effective way
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 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>0</td>
<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>
</tr>
<tr>
<td>1</td>
<td><strong>DSDA01</strong> Use predictive analytics to analyse big data and discover new relations</td>
<td><strong>DSDM01</strong> Develop and implement data strategy, in particular, Data Management Plan (DMP)</td>
<td><strong>DSENG01</strong> Use engineering principles to design, prototype data analytics applications, or develop instruments, systems</td>
<td><strong>DSRM01</strong> Create new understandings and capabilities by using scientific/research methods or similar domain related development methods</td>
</tr>
<tr>
<td>2</td>
<td><strong>DSDA02</strong> Use statistical techniques to deliver insights</td>
<td><strong>DSDM02</strong> Develop data models including metadata</td>
<td><strong>DSENG02</strong> Develop and apply computational solutions</td>
<td><strong>DSRM02</strong> Direct systematic study toward a fuller knowledge or understanding of the observable facts</td>
</tr>
<tr>
<td>3</td>
<td><strong>DSDA03</strong> Develop specialized ...</td>
<td><strong>DSDM03</strong> Collect integrate data</td>
<td><strong>DSENG03</strong> Develops specialized tools</td>
<td><strong>DSRM03</strong> Undertakes creative work</td>
</tr>
<tr>
<td>4</td>
<td><strong>DSDA04</strong> Analyze complex data</td>
<td><strong>DSDM04</strong> Maintain repository</td>
<td><strong>DSENG04</strong> Design, build, operate</td>
<td><strong>DSRM04</strong> Translate strategies into actions</td>
</tr>
<tr>
<td>5</td>
<td><strong>DSDA05</strong> Use different analytics</td>
<td><strong>DSDM05</strong> Visualise complex data</td>
<td><strong>DSENG05</strong> Secure and reliable data</td>
<td><strong>DSRM05</strong> Contribute to organizational goals</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 Social Intelligence**
  - 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
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

<table>
<thead>
<tr>
<th>Profile ID</th>
<th>Data Science Profile title</th>
<th>DSDA</th>
<th>DSDM</th>
<th>DSENG</th>
<th>DSRM</th>
<th>DSDK</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSP01</td>
<td>Data Science (group) Manager</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>DSP02</td>
<td>Data Science Infrastr Manager</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DSP03</td>
<td>Research Infrastructure Manager</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

| DSP04     | Data Scientist                              | 5    | 3    | 4     | 5    | 3    |
| DSP05     | Data Science Researcher                     | 4    | 3    | 2     | 5    | 4    |
| DSP06     | Data Science Architect                      | 4    | 3    | 5     | 3    | 3    |
| DSP07     | Data Science Applic Programmer              | 4    | 2    | 5     | 3    | 4    |
| DSP08     | Data Analyst                               | 5    | 3    | 3     | 3    | 4    |
| DSP09     | Business Analyst                            | 5    | 3    | 3     | 4    | 5    |

| DSP10     | Data Stewards                               | 3    | 5    | 3     | 3    | 3    |
| DSP11     | Digital data curator                        | 1    | 5    | 2     | 2    | 3    |
| DSP12     | Digital Librarians                          | 2    | 5    | 2     | 2    | 3    |
| DSP13     | Data Archivists                             | 1    | 5    | 1     | 1    | 3    |

| DSP14     | Large scale database designer               | 2    | 4    | 4     | 3    | 3    |
| DSP15     | Large scale database admin                  | 2    | 4    | 3     | 2    | 3    |
| DSP16     | Scientific database administrator            | 2    | 4    | 3     | 2    | 3    |

| DSP17     | Big Data facilities Operator                | 1    | 4    | 4     | 2    | 3    |
| DSP18     | Large scale data storage operator           | 1    | 4    | 3     | 1    | 1    |
| DSP19     | Scientific database operator                | 1    | 4    | 3     | 2    | 3    |

| DSP20     | Data entry/access worker                    | 2    | 1    | 2     |     |     |
| DSP21     | Data entry field workers                    | 2    | 1    | 2     |     |     |
| DSP22     | User support data services                  | 3    | 2    | 2     |     |     |

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

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

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

- Learning methods and learning models (in progress)
Outcome Based Educations and Training Model

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

• EDSF allow for customized educational courses and training modules design

Data Science Competence Framework (CF-DS) → DS Professional Profiles (DSP-P) → Learning Outcomes (LO) (OBE Learning Model) → Knowledge Units (KU) (DS-BoK) → Learning Units (LU) (MC-DS) → Tracks/Specialisations (based on DSP-P) (LO, LU, KU, courses/modules)

DS Academic Programmes → DS Professional Training

Courses → LO/KU/LU/DSP-P

Modules → LO/KU/LU/DSP-P
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>KAG2-DSENG: Data Science Engineering group including Software and Infrastructure engineering</td>
<td>Computer systems organisation for Big Data</td>
<td>Parallel and Distributed Computer Architecture</td>
<td>CCS2012: Computer systems organisation Architectures Parallel architectures</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>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>Tier 1</td>
<td>Tier 2</td>
</tr>
<tr>
<td>Software requirements and design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information theory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematical analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extensibility point for adding new courses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artificial Intelligence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge Representation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reasoning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data mining and knowledge discovery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text analysis, Data mining</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text analytics including linguistic, and struct techniques to analyze and unstructured data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine Learning algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research methodology, research cycle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modelling and experiment planning</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Mapping suggested to ACM CCS2012, DS-BoK and other related BoKs
New courses currently missing

- Data Management / Research Data Management
  - Data Curation, Data Stewardsip

- Professional issues in Data Science
  - + Ethics and responsible use of Data Science
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)
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 Developments and Actions

• Involve academic and industry experts and professional organisations to the definition of DS-BoK following from CF-DS
• Work with champion universities to practically validate the proposed EDSF
• Formally provide suggestions to ESCO for the definition of the Data Science professional profiles (occupations) family
• Formally provide suggestions for e-CF3.0 extensions for Data Science to CEN/PC 428
  – Involve national e-CF bodies and adopters where available
• Suggest required ACM CCS(2012) Classification extensions and proposal for Data Science curriculum definition
Discussion

• Questions
• Comments

• Invitation to contribution and cooperation:
  – Forum, EDISON Liaisons Groups, Champions Conference (Spring & Summer 2017)

• 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
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.
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

Domain specific components

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

Organisational roles

Cross-organisational assistive role

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
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

  - 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 (October 2016) identified need for data experts and data stewards
  - Recommendation: Allocate 5% grant funding for Data management and preservation
  - Estimation: More than 80,000 data stewards (1 per every 20 scientists)
  - Core Data Experts (as defined) need to be trained and their career perspective improved
• Clash of cultures
  – between domain specialists and e-Infrastructure specialists (i.e. IT/Computer Science)

• New data experts come from **scientific** and **engineering** cultures
  – with very different reward systems and incentives,
  – different jargons and very different skill sets.

• Consequences
  – Evident a divide between researchers and those that support research with data processing and software
    • Two communities that are both essential to Open Science have not closely co-evolved and do not converge
  – Lack of data scientists that venture out from classical computer or data science departments into other scientific fields.
Core Data Experts is a new class of colleagues with core scientific professional competencies and the communication skills to fill the gap between the two cultures.

- Core data experts are neither computer savvy research scientists nor are they hard-core data or computer scientists or software engineers.
- They should be technical data experts, though proficient enough in the content domain where they work routinely from the very beginning (experimental design, proposal writing) until the very end of the data discovery cycle.
- Converge two communities:
  - Scientists need to be educated to the point where they hire, support and respect Core Data Experts
  - Data Scientists (Core Data Experts) need to bring the value to scientific research and organisations

Implementation of the EOSC needs to include instruments to help train, retain and recognise this expertise,
- In order to support the 1.7 million scientists and over 70 million people working in innovation.
GO FAIR and IFDS

- Global Open FAIR
  - Findable – Accessible – Interoperable - Reusable
- IFDS – Internet of FAIR Data and Services = EOSC
- GO FAIR implementation approach
  - GO-BUILD
  - GO-CHANGE
  - GO-TRAIN: Training of data stewards capable of providing FAIR data services
- A critical success factor is availability of expertise in data stewardship
  - Training of a new generation of FAIR data experts is urgently needed to provide the necessary capacity.
EOSC Report Recommendations – Implementation on training and skills

• **I2.1: Set initial guiding principles to kick-start the initiative as quickly as possible.** -> **Bridge two cultures/communities**
  – A first cohort of core data experts should be trained to translate the needs for data driven science into technical specifications to be discussed with hard-core data scientists and engineers.
  – This new class of core data experts will also help translate back to the hard-core scientists the technical opportunities and limitations

• **I3: Fund a concerted effort to develop core data expertise in Europe.**
  – Substantial training initiative in Europe to locate, create, maintain and sustain the required core data expertise.
  – By 2022, to train (hundreds of thousands of) certified core data experts with a demonstrable effect on ESFRI/e-INFRA activities and prospects for long-term sustainability of this critical human resource
    • Consolidate and further develop assisting material and tools for Data Management Plans and Data Stewardship plans (including long-term preservation in FAIR status)

• **I7: Provide a clear operational timeline to deal with the early preparatory phase of the EOSC.**
  – Define training needs for the necessary data expertise and draw models for the necessary training infrastructure