EDISON Data Science Framework (EDFS):

Education and training for Data Science and data related skills

EDISON
building the data science profession

EDISON – **E**ducation for **D**ata **I**ntensive **S**cience to **O**pen **N**ew science frontiers

Grant 675419 (INFRASUPP-4-2015: CSA)

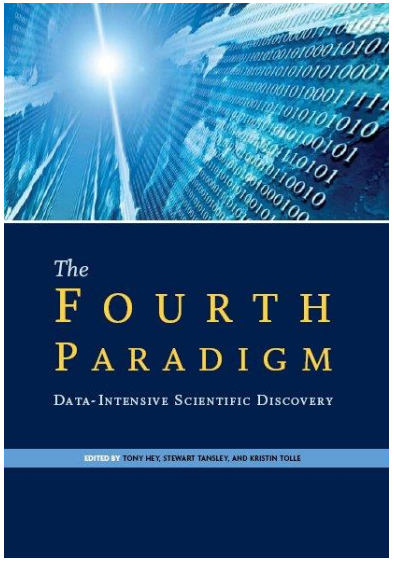
Yuri Demchenko, EDISON Project
University of Amsterdam

EDISON workshop: From Infrastructure to skills for insight and value
Amsterdam, 29 August 2017

Outline

- Background: Data driven research and demand for new skills
  - Recent reports, studies and facts
- EDISON Data Science Framework (EDSF)
  - Data Science competences and skills
  - Essential Data Scientist professional skills: Thinking and doing like Data Scientist
- Data Science Professional Profiles
- Data Science Body of Knowledge and Model Curriculum
- Use of EDSF and Example curricula
  - Competences assessment
  - Building Data Science team
- Roadmap recommendations
- Discussion and Questions

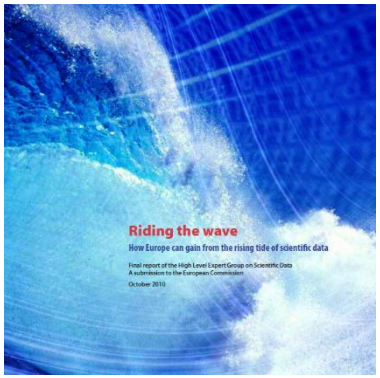
Visionaries and Drivers: Seminal works, High level reports, Activities



The Fourth Paradigm: Data-Intensive Scientific Discovery.

By Jim Gray, Microsoft, 2009. Edited by Tony Hey, Kristin Tolle, et al.

http://research.microsoft.com/en-us/collaboration/fourthparadigm/



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

Final report of the High Level Expert Group on Scientific Data. October 2010.

http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

RDA RESEARCH DATA ALLIANCE

Research Data Sharing without barriers

https://www.rd-alliance.org/

The Data Harvest: How sharing research data can yield knowledge, jobs and growth.

An RDA Europe Report. December 2014

https://rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html

HLEG report on European Open Science Cloud

(October 2016)

Emergence of Cognitive Technologies

(IBM Watson, Cortana and others)

Riding the wave (2010): How Europe can gain from the rising tide of scientific data.

- "Unlocking the full value of scientific data"
  - Neelie Kroes, *Vice-President of the European Commission, responsible for the Digital Agenda*
- Just how students will be trained in the future, or how the **profession of "data scientist"** will be developed, are among the questions the resolution of which is still evolving and will present intellectual challenges for both privately and publicly supported research.
  - John Wood, HLEG Chair
- Vision 2030: "Our vision is a scientific e-Infrastructure that supports seamless access, use, re-use and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure."
- Proposed set of actions
  - **4. Train a new generation of data scientists, and broaden public understanding**
    We urge that the European Commission promote, and the member-states adopt, new policies to foster the development of advanced-degree programmes at our major universities for the emerging field of data scientist. We also urge the member-states to include data management and governance considerations in the curricula of their secondary schools, as part of the IT familiarisation programmes that are becoming common in European education.

The Data Harvest (2014): How sharing research data can yield knowledge, jobs and growth

- Planning the data harvest – John Wood
- The era of data driven science
- We want the right minds, with the right data, at the right time. That's a tall order that requires change in:
  - The way science works and scientists think
  - How scientific institutions operate and interact
  - How scientists are trained and employed

Recommendation 2

- DO promote data literacy across society, from researcher to citizen. Embracing these new possibilities requires training and cultural education – inside and outside universities. Data science must be promoted
  - A first-class science: Data sharing provides the foundation for a new branch of science.
  - Data education: Training in the use, evaluation and responsible management of data needs to be embedded in curricula, across all subjects, from primary school to university.
  - Training within EU projects
  - Government and public sector training

HLEG EOSC Report Essentials – **Core Data Experts**

- **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.

EOSC Report Recommendations – Implementation on training and skills

- **I2.1: Set initial guiding principles to kick-start the initiative as quickly as possible.**
  - 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**

Initiatives: GO FAIR and IFDS

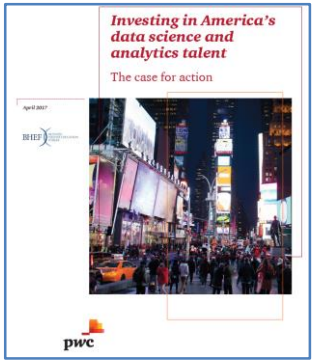
- Global Open FAIR
  - Findable – Accessible – Interoperable - Reusable
- IFDS – Internet of FAIR Data and Services = EOSC
- GO FAIR implementation approach
  - GO-BUILD
  - GO-CHANGE
  - GO-TRAIN: Training of data stewards capable of providing FAIR data services
- A critical success factor is availability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

Industry reports on Data Science Analytics and Data enabled skills demand

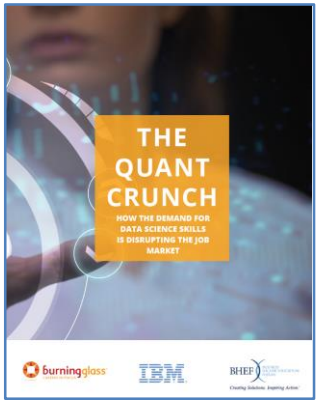
- Final Report on European Data Market Study by IDC (Feb 2017)
  - The EU data market in 2016 estimated EUR 60 Bln (growth 9.5% from EUR 54.3 Bln in 2015)
    - Estimated EUR 106 Bln in 2020
  - Number of data workers 6.1 mln (2016) - increase 2.6% from 2015
    - Estimated EUR 10.4 million in 2020
  - Average number of data workers per company 9.5 - increase 4.4%
  - Gap between demand and supply estimated 769,000 (2020) or 9.8%

- PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017)
  - http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent
  - 2.35 mln postings, 23% Data Scientist, 67% DSA enabled jobs
  - DSA enabled jobs growing at higher rate than main Data Science jobs

- Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017)
  - https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF
  - DSA enabled jobs takes 45-58 days to fill: 5 days longer than average
  - Commonly required work experience 3-5 yrs

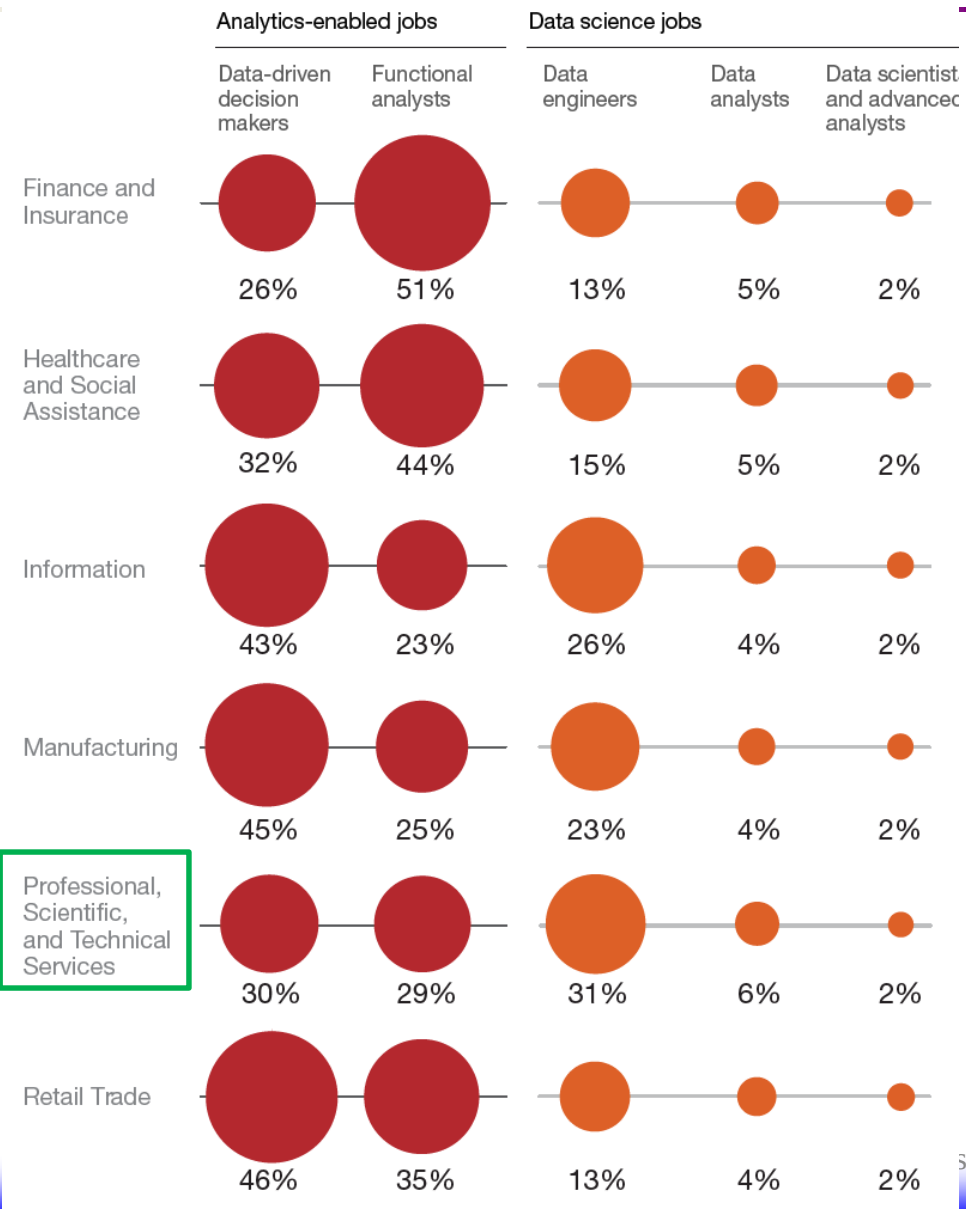


Citing EDISON and EDSF



Influenced by EDISON

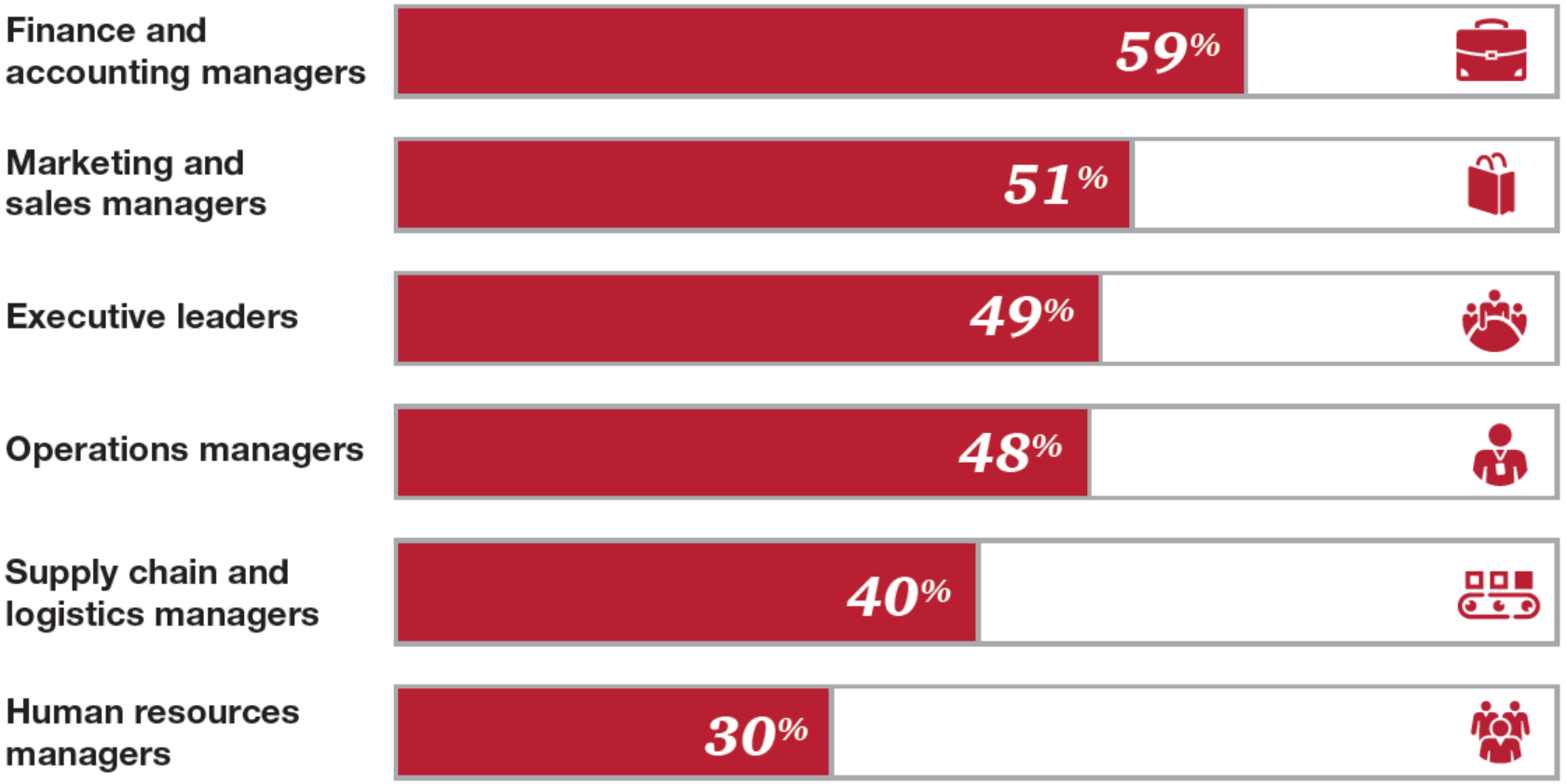
PwC&BHEF: Demand for DSA enabled jobs



Demand for business people with analytics skills, not just data scientists

- Of 2.35 million job postings in the US
  - 23% Data Scientist
  - 67% DSA enabled jobs
- Strong demand for managers and decision makers with Data Science (data analytics) skills/understanding
  - Challenge to deliver actionable knowledge and competences to CEO level managers

PwC&BHEF: Data Science and Data Analytics Competences for Managers and Decision Makers

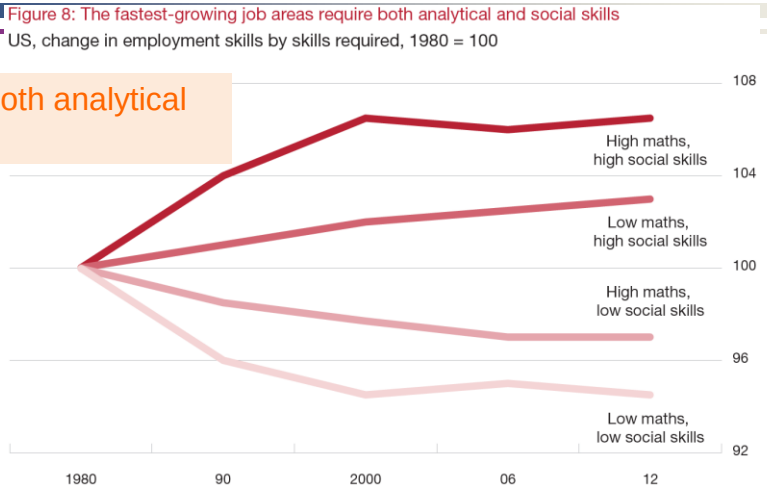


Percent of employers who say data science and analytics skills will be 'required of all managers' by 2020

- Source: BHEF and Gallup, *Data Science and Analytics Business Survey* (December 2016).

PwC&BHEF: Skills that are tough to find

Faster growing jobs require both analytical and social skills



97% Cybersecurity

95% Data science and analytics

83% Critical thinking and problem solving

79% Design/ systems thinking

79% Innovation and creativity

78% Global perspective

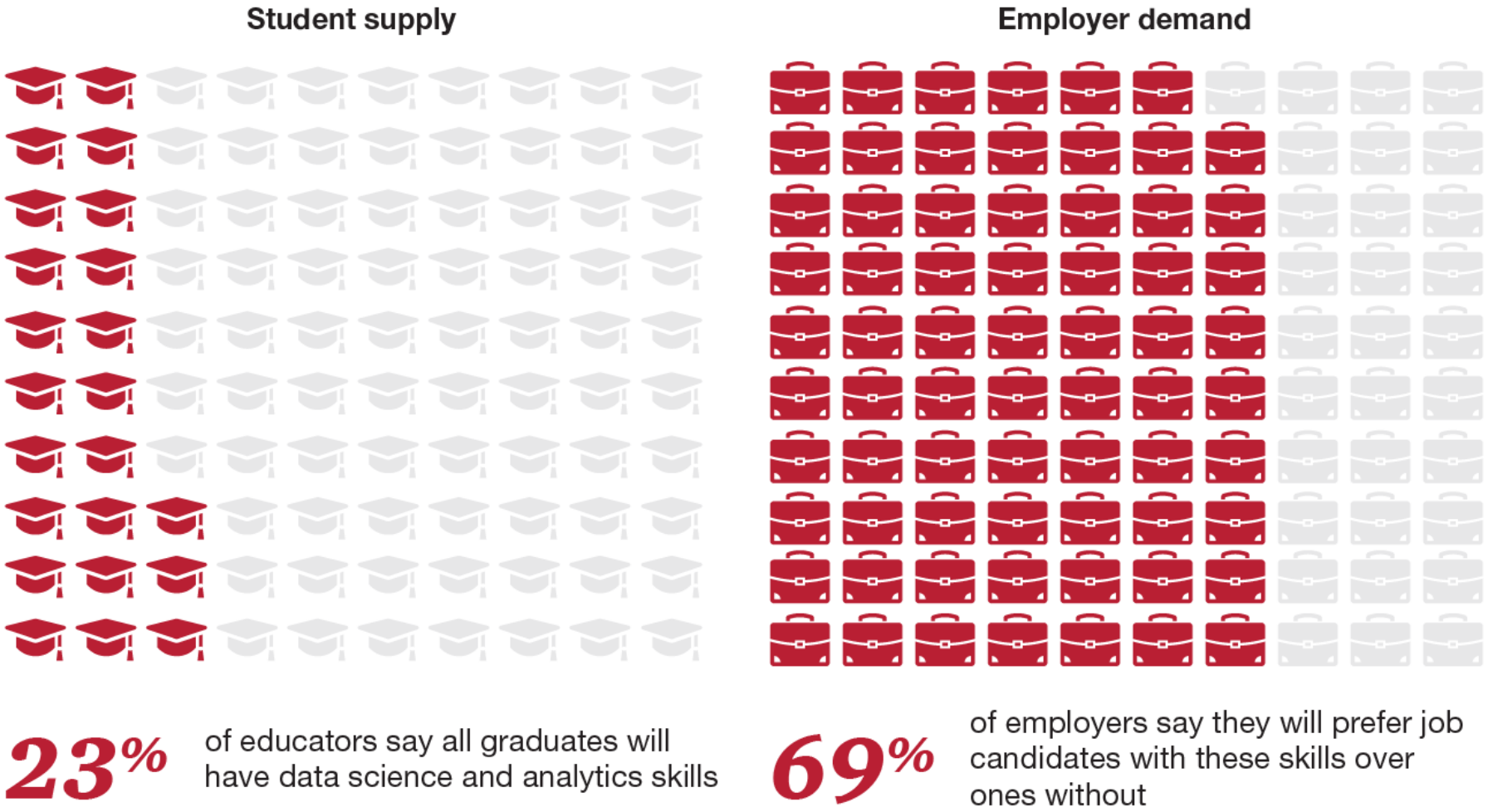
78% Cognitive flexibility

74% Cross-disciplinary ability

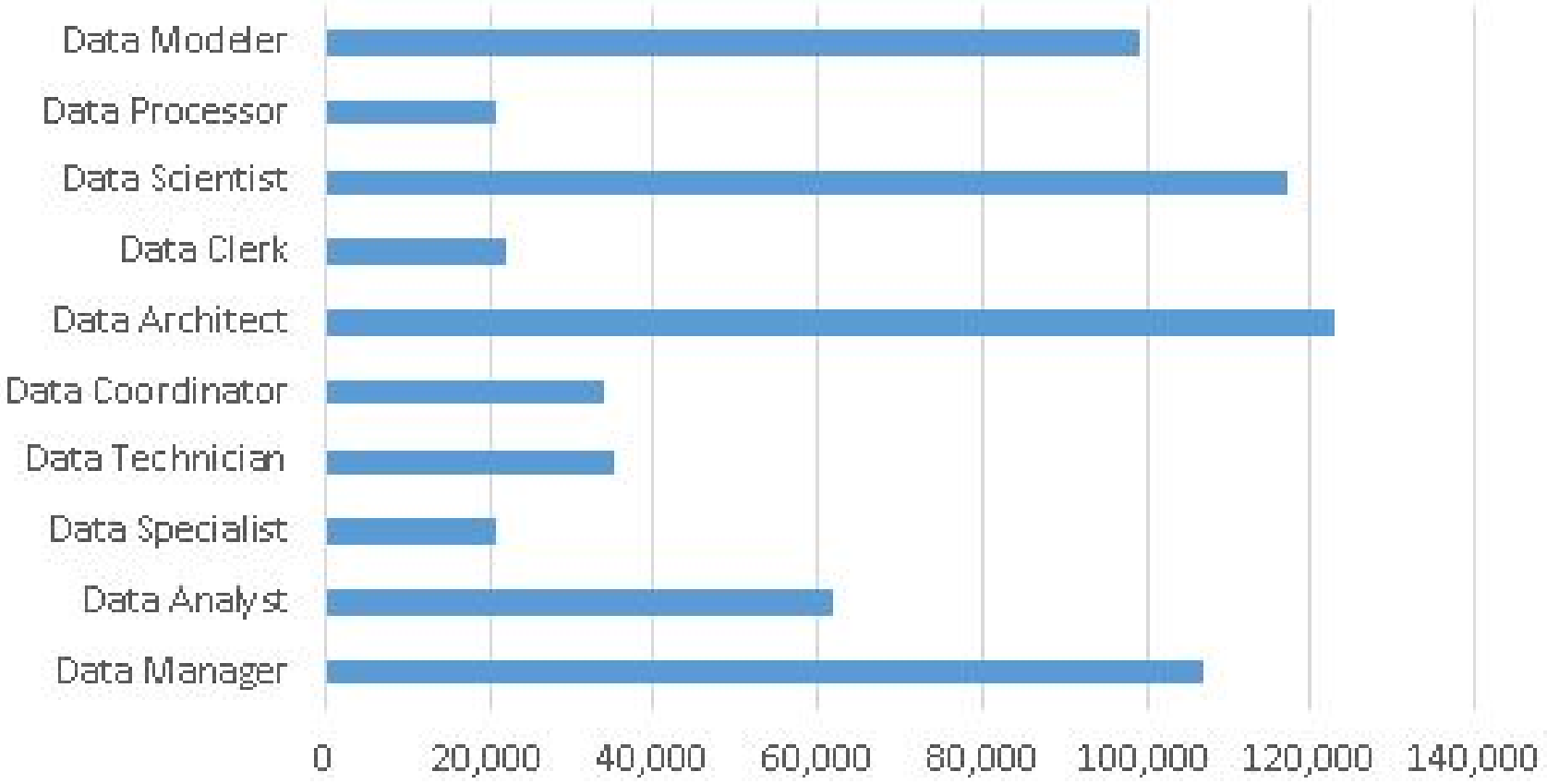
To be mapped to Competences, Knowledge, Skills and Personal (soft) Skills

Source: Business Roundtable (2017).

PwC&BHEF: Data Science and Analytics skills, by 2021: The supply-demand challenge



Closer look at Data related Jobs and Salaries (2016)



Source: The Jon Market for Data Professionals, by Robert R Downs, SciDataCon2016
http://www.scidatacon.org/2016/sessions/98/poster/51/

2017: What is the answer from RI and Science?

- New study? New report focused on Scientific Data Infrastructure?
- Who is the Rider? Who is the Harvester?

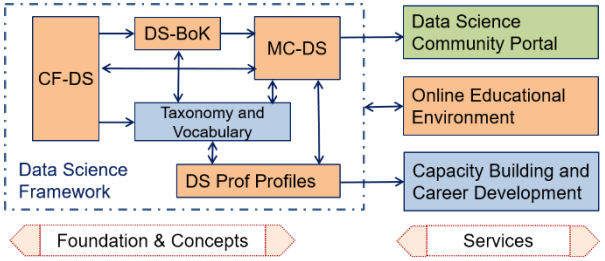
- Demand for new data driven competences and skills and new type of professionals – Core data experts
  - The whole family of professions is required

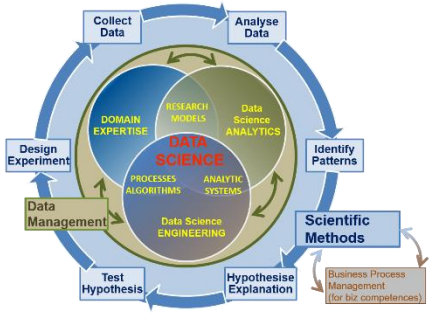
Challenge for Education: Sustainable ICT and Data Skills Development

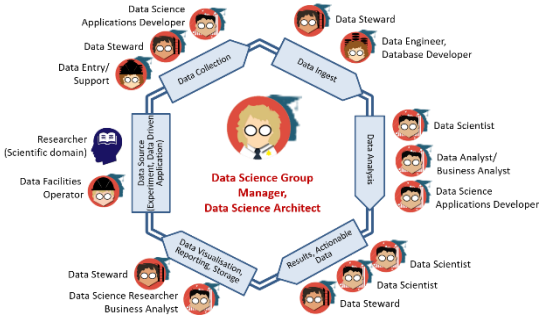
- Educate vs Train
  - Training is a short term solution
  - Education is a basis for sustainable skills development
- Technology focus changes every 3-4 years
  - Study: 50% of academic curricula are outdated at the time of graduation
- Lack of necessary skills leads to *underperforming projects* and organisations and *loose of competitiveness*
  - Challenge: Policy and decision makers still don't include planning human factor (competences and skills) as a part of the technology strategy
- Need to change the whole skills management paradigm
  - **Dynamic (self-) re-skilling:** Continuous professional development and **shared responsibility between employer and employee**
  - Professional and workplace skills and career management as a part of professional orientation
- Millennials factor and changing nature of workforce

EDISON Products for Data Science Skills Management and Tailored Education

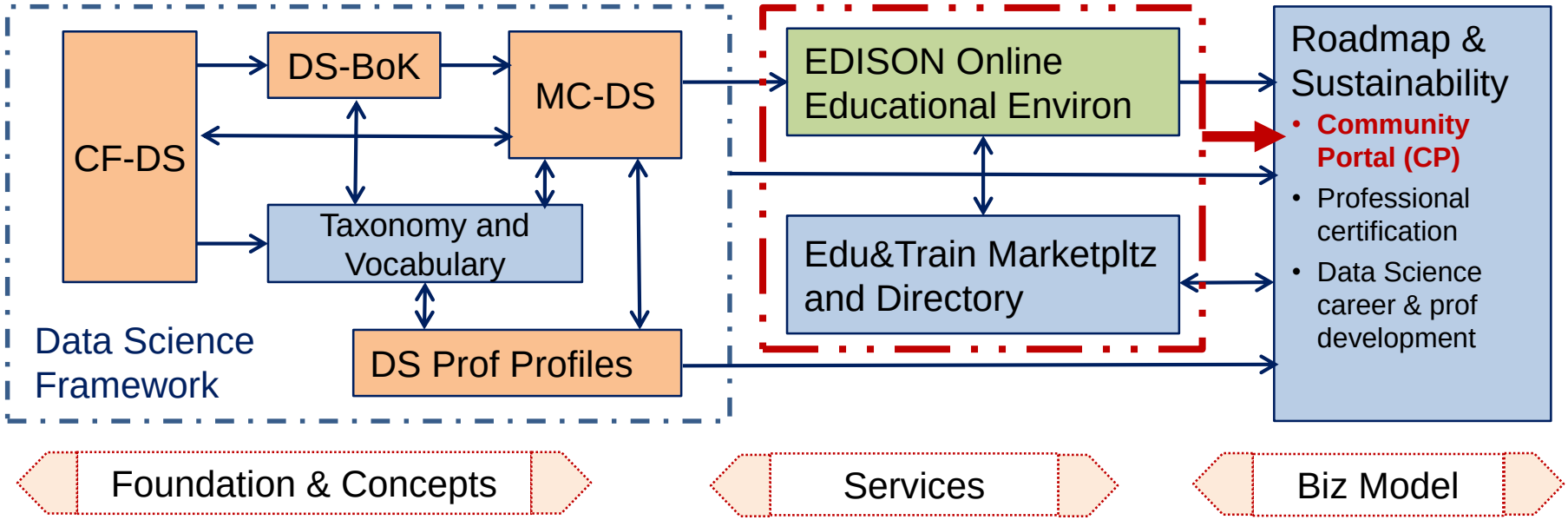
- EDISON Data Science Framework (EDSF)
  - Compliant with EU standards on competences and professional occupations e-CFv3.0, ESCO
  - Customisable courses design for targeted education and training
- Skills development and career management for Core Data Experts and related data handling professions
- Capacity building and Data Science team design
- Academic programmes and professional training courses (self) assessment and design
- EU network of Champion universities pioneering Data Science academic programmes
- Engagement in relevant RDA activities and groups
- Cooperation with International professional organisations IEEE, ACM, BHEF, APEC (AP Economic Cooperation)







EDISON Data Science Framework (EDSF)



EDISON Framework components

- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- DSP – Data Science Professional profiles
- Data Science Taxonomies and Scientific Disciplines Classification
- EOEE - EDISON Online Education Environment

Methodology

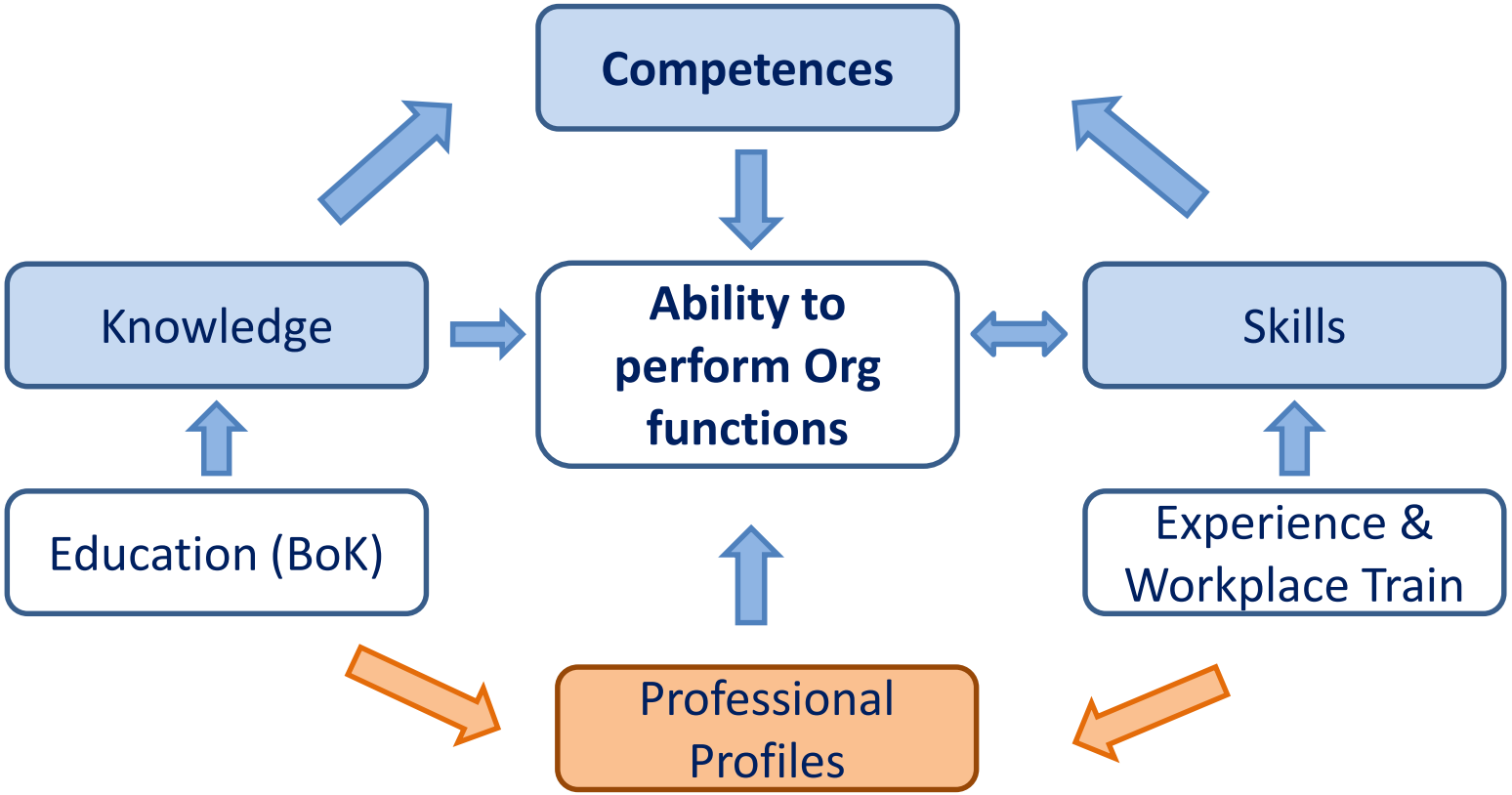
- ESDF development based on job market study, existing practices in academic, research and industry.
- Review and feedback from the ELG, expert community, domain experts.
- Input from the champion universities and community of practice.

What challenges related to skills management the EDSF can help to address?

1. Guide researchers in using right methods and tools, latest Data Analytics technologies to extracting value from scientific data
2. Educate and train RI engineers dev to build modern data intensive research infrastructure and understand trends and project for future
3. Develop new data analytics tools and ensure continuous improvement (agile model, DevOps)
4. Correctly organise and manage data, make them accessible (adhering FAIR principles), education new profession of Data Stewards
5. Help managers to facilitate career dev for researchers and organise effective teams
6. Ensure skills and expertise sustain in organisation
7. Help research institutions to sustain in competition with industry and business in data science talent hunting

Competences Map to Knowledge and Skills

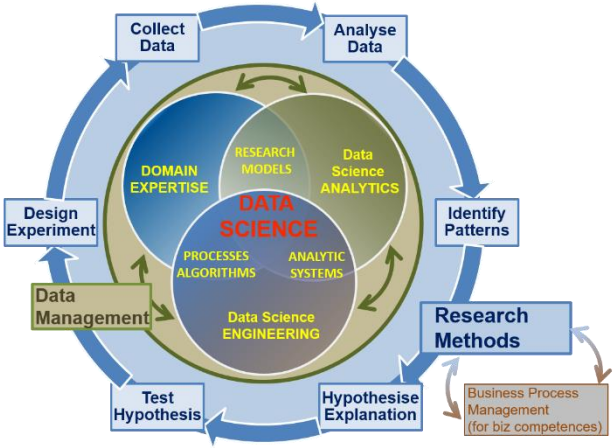
- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results



Data Scientist definition

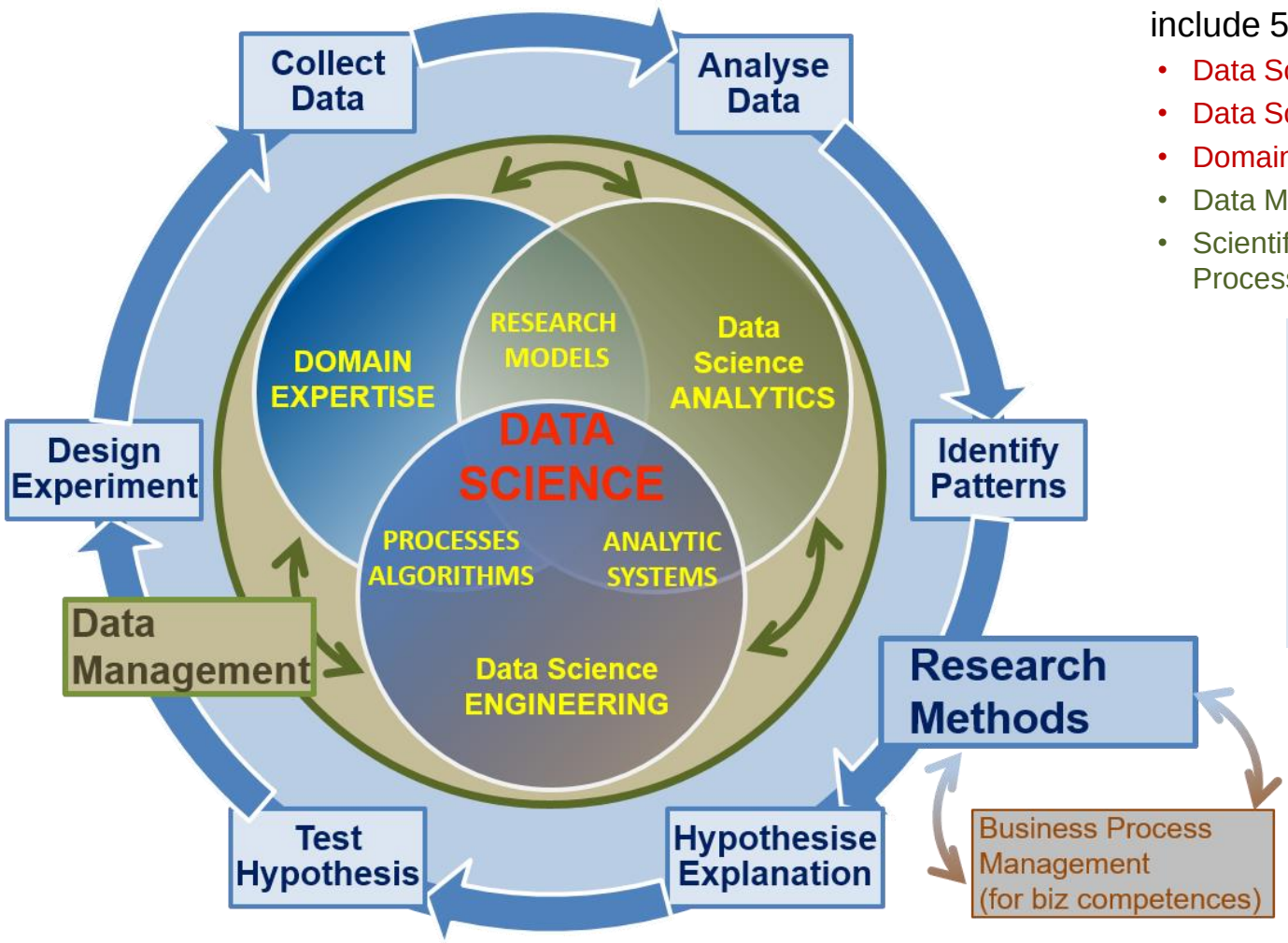
Based on the definitions by NIST SP1500 – 2015, extended by EDISON

- *A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in **business needs, domain knowledge, analytical skills, and programming and systems engineering expertise** to manage the end-to-end scientific method process through each stage in the **big data lifecycle** till the delivery of an **expected scientific and business value** to organisation or project.*



- Core Data Science competences and skills groups
  - **Data Science Analytics** (including Statistical Analysis, Machine Learning, Business Analytics)
  - **Data Science Engineering** (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
  - **Domain Knowledge and Expertise** (Subject/Scientific domain related)
- EDISON identified 2 additional competence groups demanded by organisations
  - **Data Management, Data Governance, Stewardship, Curation, Preservation**
  - **Research Methods and/vs Business Processes/Operations**
- **Data Science professional skills**: Thinking and acting like Data Scientist – required to successfully develop as a Data Scientist and work in Data Science teams

Data Science Competence Groups - Research



Data Science Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Scientific Methods or Business Process Management

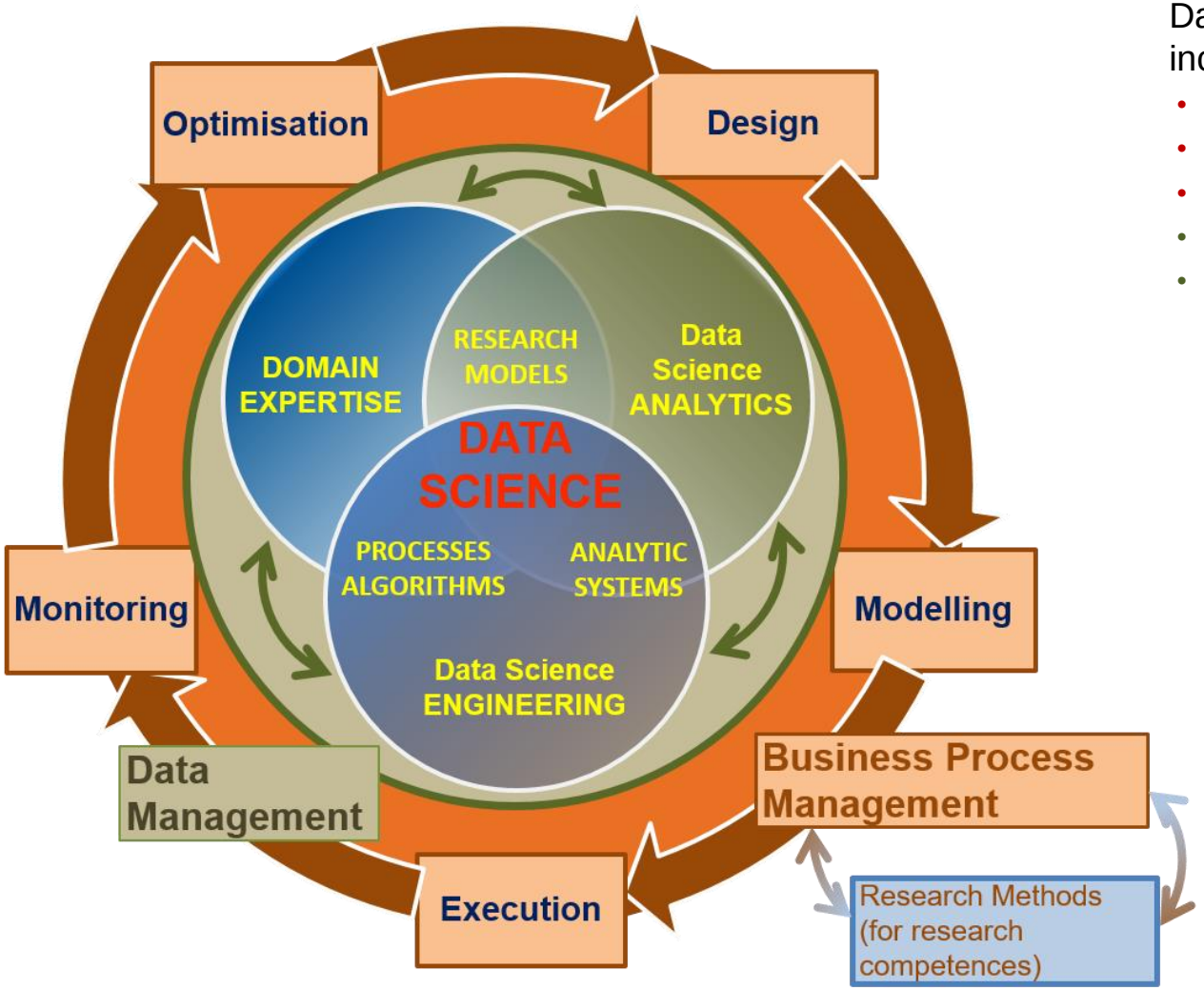
Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesis Explanation
- Test Hypothesis

Business Operations

- Operations Strategy
- Plan
- Design & Deploy
- Monitor & Control
- Improve & Re-design

Data Science Competences Groups – Business



Data Science Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Scientific Methods or Business Process Management

Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

Business Process Operations/Stages

- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design

Identified Data Science Competence Groups

	Data Science Analytics (DSDA)	Data Science Engineering (DSENG)	Data Management and Governance (DSDM)	Research/Scientific Methods and Project Management (DSRMP)	Data Science Domain Knowledge, e.g. Business Analytics (DSDK/DSBPM)
0	Use appropriate data analytics and statistical techniques on available data to deliver insights into research problem or org. processes and support decision making	Use engineering principles and modern computer technology to research, design, implement new data analytics applications, develop experiments, processes, instruments, systems and infrastructures to support data handling during the whole data lifecycle	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	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	DSDK/DSBA Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
1	**DSDA01** Effectively use variety of data analytics techniques	**DSENG01** Use engineering principles (general and software) to research, design, develop and implement new instruments and applications	DSDM01 Develop and implement data strategy, in particular, Data Management Plan (DMP)	**DSRMP01** Create new understandings and capabilities by using scientific/ research methods	**DSBPM01** Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
2	**DSDA02** Apply designated quantitative techniques	**DSENG02** Develop and apply computer methods to domain related problems	DSDM02 Develop data models including metadata	**DSRMP02** Direct systematic study toward a fuller knowledge or understanding of the observable facts	**DSBPM02** Participate strategically and tactically in financial decisions
3	**DSDA03** Pull together data from diff sources …	**DSENG03** Develop and prototype data analytics applications	DSDM03 Collect integrate data	**DSRMP03** Undertakes creative work	**DSBPM03** Provides support services to other
4	**DSDA04** Use diff perform techniques	**DSENG04** Develop, deploy operate Big Data storage	DSDM04 Maintain repository	**DSRMP04** Translate strategies into actions	**DSBPM04** Analyse data for marketing
5	**DSDA05** Develop analytics applic	**DSENG05** Apply security mechanisms	DSDM05 Visualise cmplx data	**DSRMP05** Contribute to organis goals	**DSBPM05** Analyse optimise customer relatio
6	**DSDA06** Visualise results of analysis, dashboards	**DSENG06** Design, build, operate SQL and NoSQL	**DSRM06** Develop and manage prolicies	**DSRMP06** Develop and guide data driven projects	**DSBPM06** Analyse data for marketing

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: Data analytics 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

Key Data Science Analytics Competences by EDISON and DARE Project for APEC countries

- Data Science Analytics, Data Science Engineering, Data Management, Research Methods, Domain Knowledge
  - 3 levels of proficiency defined: Entry/Associate, Professional, Expert/Lead
- **Foundational competences** (starting from entry level to expert level)
  - Statistics, Probability theory, mathematics, calculus
  - Statistical programming languages, frameworks, tools
  - Computational methods and document processing tools (including Excel, Office visualization, or similar)
  - Data Visualisation and tools
- **21st Century Skills** and **Data Science Professional skills**

Data Science Professional Skills: Thinking and Acting like Data Scientist

1. **Recognise value of data**, work with raw data, exercise good data intuition, use SN and open data
2. Accept (be ready for) **iterative development**, know when to stop, comfortable with failure, accept the symmetry of outcome (both positive and negative results are valuable)
3. Good **sense of metrics**, understand importance of the results validation, never stop looking at individual examples
4. **Ask the right questions**
5. **Respect domain/subject matter knowledge** in the area of data science
6. **Data driven problem solver** and **impact-driven mindset**
7. **Be aware about power and limitations** of the main machine learning and data analytics algorithms and tools
8. Understand that most of **data analytics algorithms are statistics and probability based**, so any answer or solution has some degree of probability and represent an optimal solution for a number variables and factors
9. Recognise what things are **important** and what things are **not important** (in data modeling)
10. Working in **agile environment** and coordinate with other roles and team members
11. Work in **multi-disciplinary team**, ability to communicate with the domain and subject matter experts
12. Embrace **online learning**, continuously improve your knowledge, use **professional netw**orks and communities
13. **Story Telling**: Deliver actionable result of your analysis
14. **Attitude**: Creativity, curiosity (willingness to challenge status quo), commitment in finding new knowledge and progress to completion
15. **Ethics and responsible use** of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)

Data Science Professional Skills: Thinking and Acting like Data Scientist (1)

1. **Recognise value of data**, work with raw data, exercise good data intuition, use SN and Open Data
2. Accept (be ready for) **iterative development**, know when to stop, comfortable with failure, accept the symmetry of outcome (both positive and negative results are valuable)
3. Good **sense of metrics**, understand importance of the results validation, never stop looking at individual examples
4. **Ask the right questions**
5. **Respect domain/subject matter knowledge** in the area of data science
6. **Data driven problem solver** and **impact-driven mindset**
7. **Be aware about power and limitations** of the main machine learning and data analytics algorithms and tools
8. Understand that most of **data analytics algorithms are statistics and probability based**, so any answer or solution has some degree of probability and represent an optimal solution for a number variables and factors

Data Science Professional Skills: Thinking and Acting like Data Scientist (2)

9. Recognise what things are **important** and what things are **not important** (in data modeling)
10. Working in **agile environment** and coordinate with other roles and team members
11. Work in **multi-disciplinary team**, ability to communicate with the domain and subject matter experts
12. Embrace **online learning**, continuously improve your knowledge, use **professional netw**orks and communities
13. **Story Telling**: Deliver actionable result of your analysis
14. **Attitude**: Creativity, curiosity (willingness to challenge status quo), commitment in finding new knowledge and progress to completion
15. **Ethics and responsible use** of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)

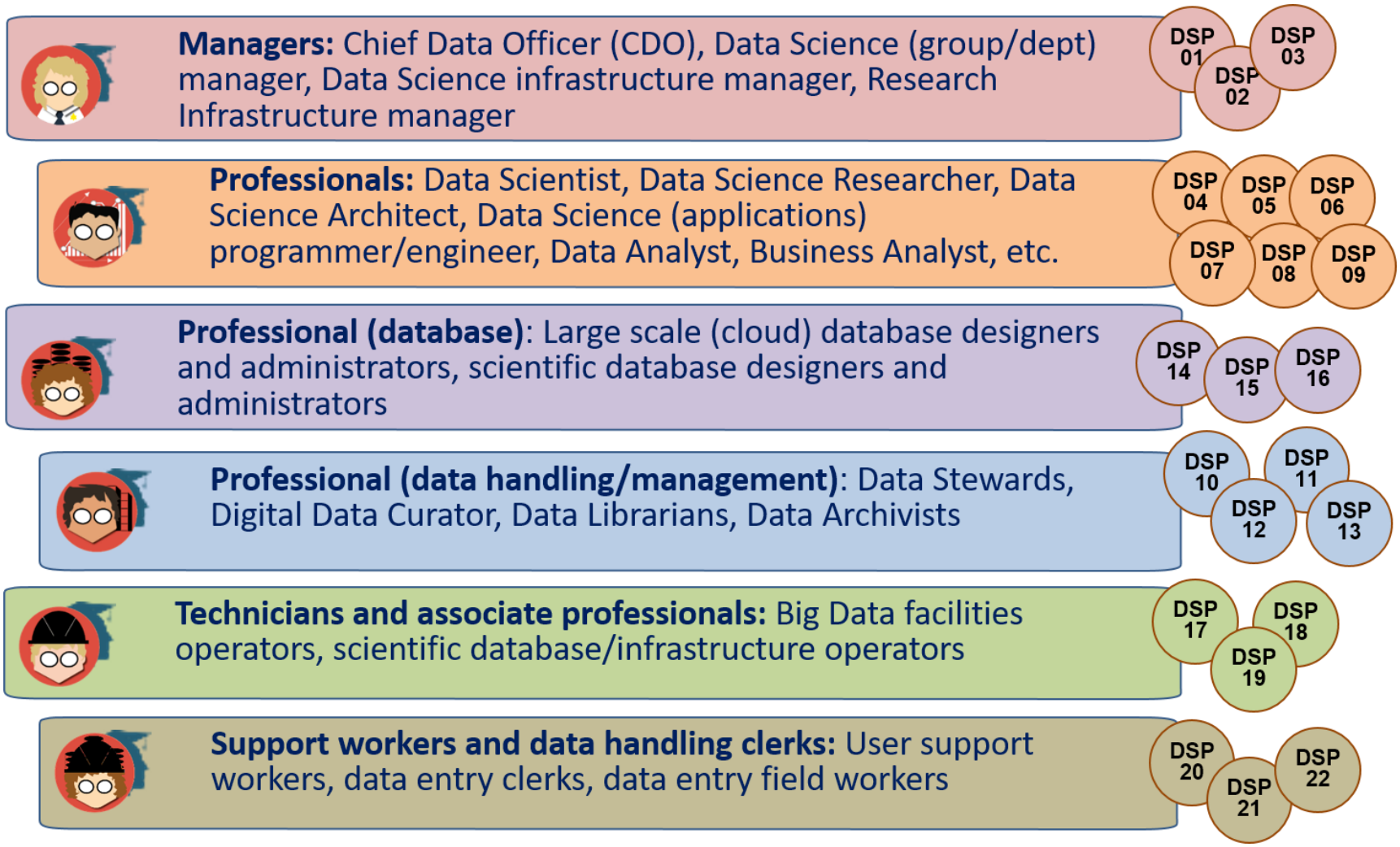
21st Century Skills (DARE & BHEF & EDISON)

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

Practical Application of the CF-DS

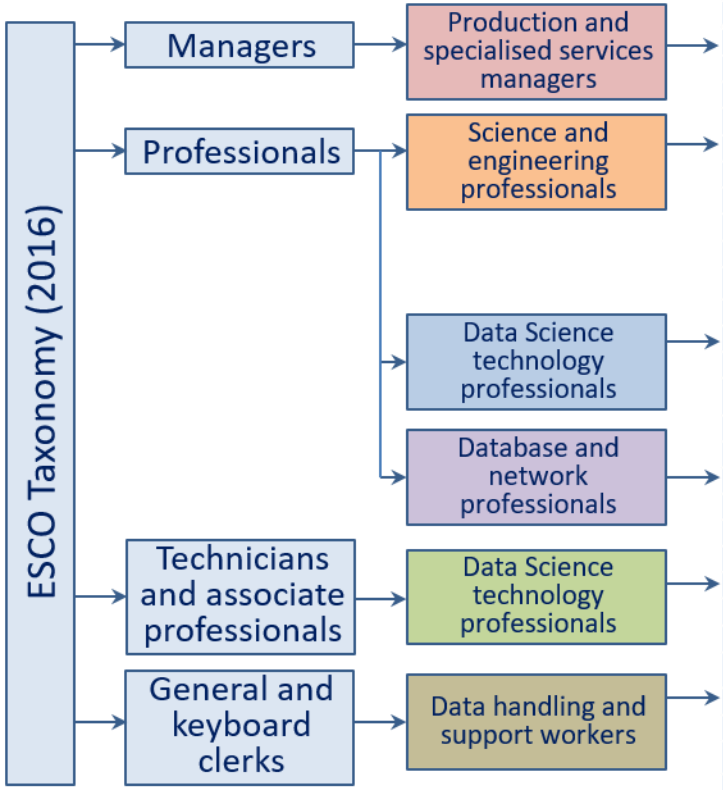
- Basis for the definition of the Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS)
  - CF-DS => Learning Outcomes (MC-DS) => Knowledge Areas (DS-BoK)
  - CF-DS => Data Science taxonomy of scientific subjects and vocabulary
- Data Science professional profiles definition
  - 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

Data Science Professions Family – By Groups



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

DSP Profiles mapping to ESCO Taxonomy High Level Groups



Profile ID	Data Science Profile title	DSDA	DSDM	DSENG	DSRM	DSDK
Data Science Services/Infrastructure Managers						
DSP01	Data Science (group) Manager	3	4	3	3	2
DSP02	Data Science Infrastr Manager	2	4	4	2	2
DSP03	Research Infrastructure Manager	2	4	4	3	2
Data Scince Professionals						
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
Data handling professionals not elsewhere classified						
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
Database and network professionals not elsewhere classified						
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
Data Infrastructure engineers and technicians						
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
Data and information entry and access						
DSP20	Data entry/access worker		2	1		2
DSP21	Data entry field workers		2	1		2
DSP22	User support data services		3	2		2

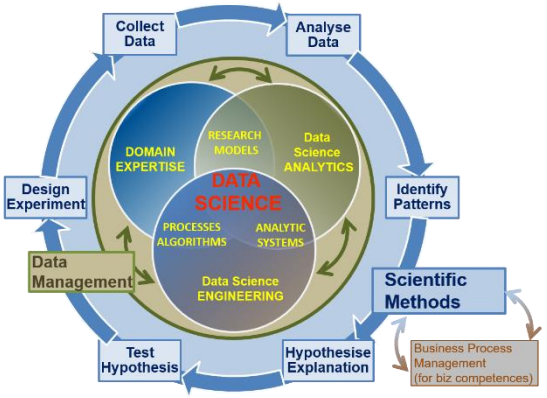
- DSP Profiles mapping to corresponding CF-DS Competence Groups
  - Relevance level from 5 – maximum to 1 – minimum

EDSF for 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

Data Science Body of Knowledge (DS-BoK)

DS-BoK Knowledge Area Groups (KAG)



- KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRM: *Research Methods and Project Management group*
- KAG5-DSBA: Business Analytics and Business Intelligence

- KAG* - DSDK: Data Science domain knowledge to be defined by related expert groups

Data Science Body of Knowledge (1)

KA Groups	Suggested DS Knowledge Areas (KA)	Knowledge Areas from existing BoK and CCS2012 scientific subject groups
KAG1-DSDA: Data Science Analytics	KA01.01 (DSDA.01/SMDA) Statistical methods for data analysis KA01.02 (DSDA.02/ML) Machine Learning KA01.03 (DSDA.03/DM) Data Mining KA01.04 (DSDA.04/TDM) Text Data Mining KA01.05 (DSDA.05/PA) Predictive Analytics KA01.06 (DSDA.06/MODSIM) Computational modelling, simulation and optimisation	There is no formal BoK defined for Data Analytics. Data Science Analytics related scientific subjects from CCS2012: CCS2012: Computing methodologies CCS2012: Mathematics of computing CCS2012: Computing methodologies
KAG2-DSENG: Data Science Engineering	KA02.01 (DSENG.01/BDI) Big Data Infrastructure and Technologies KA02.02 (DSENG.02/DSIAPP) Infrastructure and platforms for Data Science applications KA02.03 (DSENG.03/CCT) Cloud Computing technologies for Big Data and Data Analytics KA02.04 (DSENG.04/SEC) Data and Applications security KA02.05 (DSENG.05/BDSE) Big Data systems organisation and engineering KA02.06 (DSENG.06/DSAPPD) Data Science (Big Data) applications design KA02.07 (DSENG.07/IS) Information systems (to support data driven decision making)	ACM CS-BoK selected KAs: AR - Architecture and Organization (including computer architectures and network architectures) CN - Computational Science IM - Information Management SE - Software Engineering (can be extended with specific SWEBOK KAs) SWEBOK selected KAs • Software requirements • Software design • Software engineering process • Software engineering models and methods • Software quality Data Science Analytics related scientific subjects from CCS2012

Data Science Body of Knowledge (2)

KA Groups	Suggested DS Knowledge Areas (KA)	Knowledge Areas from existing BoK and CCS2012 scientific subject groups
KAG3-DSDM: Data Management	KA03.01 (DSDM.01/DMORG) General principles and concepts in Data Management and organisation KA03.02 (DSDM.02/DMS) Data management systems KA03.03 (DSDM.03/EDMI) Data Management and Enterprise data infrastructure KA03.04 (DSDM.04/DGOV) Data Governance KA03.05 (DSDM.05/BDST0R) Big Data storage (large scale) KA03.06 (DSDM.05/DLIB) Digital libraries and archives	DM-BoK selected KAs (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, and (11) Data Quality.
KAG4-DSRM: Research Methods and Project Management	KA04.01 (DSRMP.01/RM) Research Methods KA04.01 (DSRMP.02/PM) Project Management	There are no formally defined BoK for research methods PMI-BoK selected KAs • Project Integration Management • Project Scope Management • Project Quality • Project Risk Management
KAG5-DSBPM: Business Analytics	KA05.01 (DSBA.01/BAF) Business Analytics Foundation KA05.02 (DSBA.02/BAEM) Business Analytics organisation and enterprise management	BABOK selected KAs *) Business Analysis Planning and Monitoring Requirements Life Cycle Management Solution Evaluation and improvements recommendation

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
  - Knowledge levels: Familiarity, Usage, Assessment (based in Bloom's Taxonomy)
- 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)

Knowledge levels for Learning Outcomes

Level	Action Verbs
Familiarity	Choose, Classify, Collect, Compare, Configure, Contrast, Define, Demonstrate, Describe, Execute, Explain, Find, Identify, Illustrate, Label, List, Match, Name, Omit, Operate, Outline, Recall, Rephrase, Show, Summarize, Tell, Translate
Usage	Apply, Analyze, Build, Construct, Develop, Examine, Experiment with, Identify, Infer, Inspect, Model, Motivate, Organize, Select, Simplify, Solve, Survey, Test for, Visualize
Assessment	Adapt, Assess, Change, Combine, Compile, Compose, Conclude, Criticize, Create, Decide, Deduct, Defend, Design, Discuss, Determine, Disprove, Evaluate, Imagine, Improve, Influence, Invent, Judge, Justify, Optimize, Plan, Predict, Prioritize, Prove, Rate, Recommend, Solve

Data Science Data Analytics (KAG1 – DSDA) related courses

- KA01.01 (DSDA/SMDA) Statistical methods, including Descriptive statistics, exploratory data analysis (EDA) focused on discovering new features in the data, and confirmatory data analysis (CDA) dealing with validating formulated hypotheses;
- KA01.02 (DSDA/ML) Machine learning and related methods for information search, image recognition, decision support, classification;
- KA01.03 (DSDA/DM) Data mining is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes;
- KA01.04 (DSDA/TDM) Text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data;
- KA01.05 (DSDA/PA) Predictive analytics focuses on application of statistical models for predictive forecasting or classification;
- KA01.06 (DSDA/MODSIM) Computational modelling, simulation and optimisation.

Data Science Engineering (KAG2-DSENG)

- KA02.01 (DSENG/BDI) Big Data infrastructure and technologies, including NOSQL databased, platforms for Big Data deployment and technologies for large-scale storage;
- KA02.02 (DSENG/DSIAPP) Infrastructure and platforms for Data Science applications, including typical frameworks such as Spark and Hadoop, data processing models and consideration of common data inputs at scale;
- KA02.03 (DSENG/CCT) Cloud Computing technologies for Big Data and Data Analytics;
- KA02.04 (DSENG/SEC) Data and Applications security, accountability, certification, and compliance;
- KA02.05 (DSENG/BDSE) Big Data systems organization and engineering, including approached to big data analysis and common MapReduce algorithms;
- KA02.06 (DSENG/DSAPPD) Data Science (Big Data) application design, including languages for big data (Python, R), tools and models for data presentation and visualization;
- KA02.07 (DSENG/IS) Information Systems, to support data-driven decision making, with focus on data warehouse and data centers.

KAG3-DSDM: *Data Management group: data curation, preservation and data infrastructure*

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

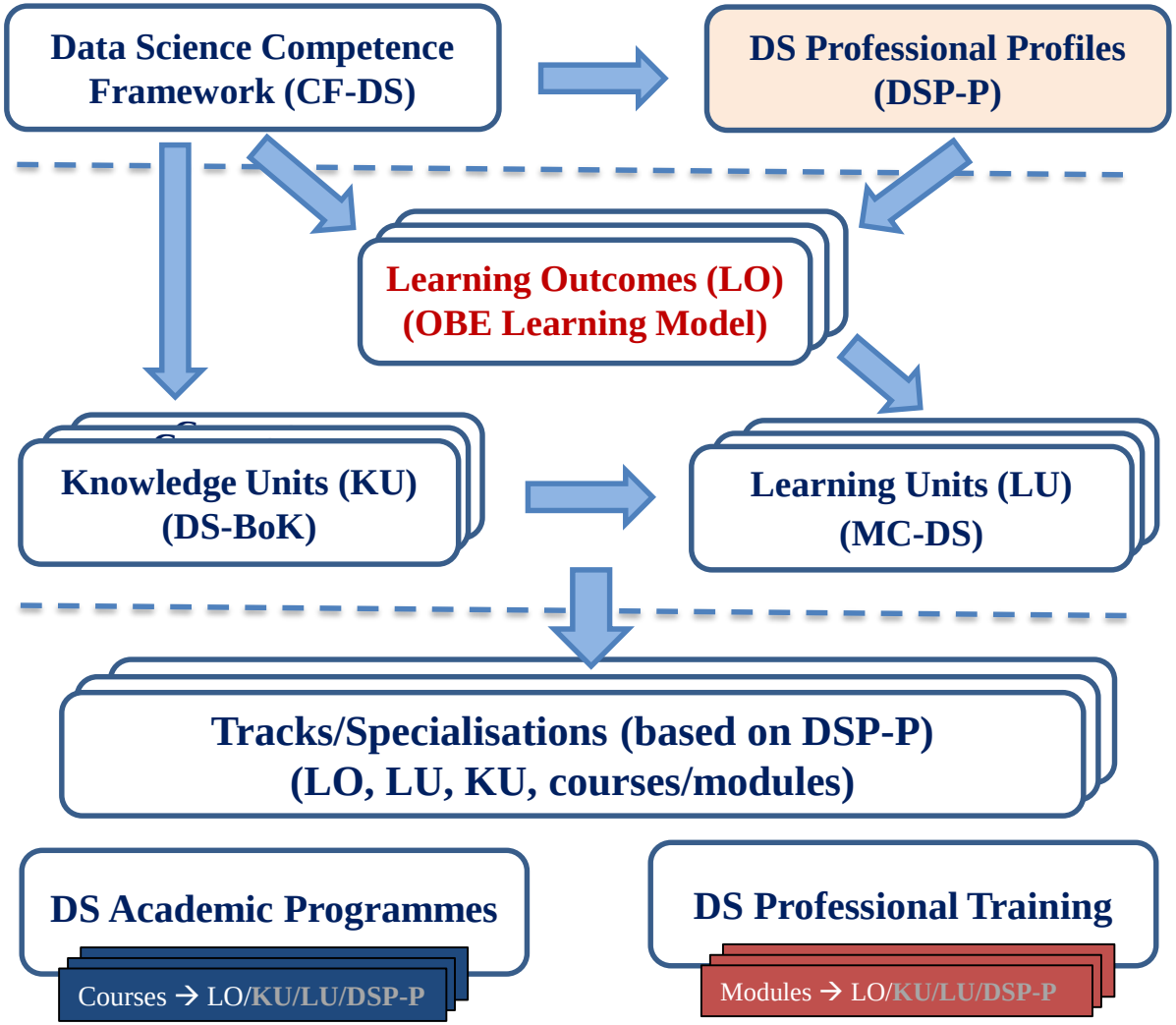
(13) Data Management Plan

(14) Open Science, Open Data, Open Access, ORCID

(15) Responsible data use

- Highlighted in red: Considered (Research) Data Management literacy (minimum required knowledge)

Outcome Based Educations and Training Model



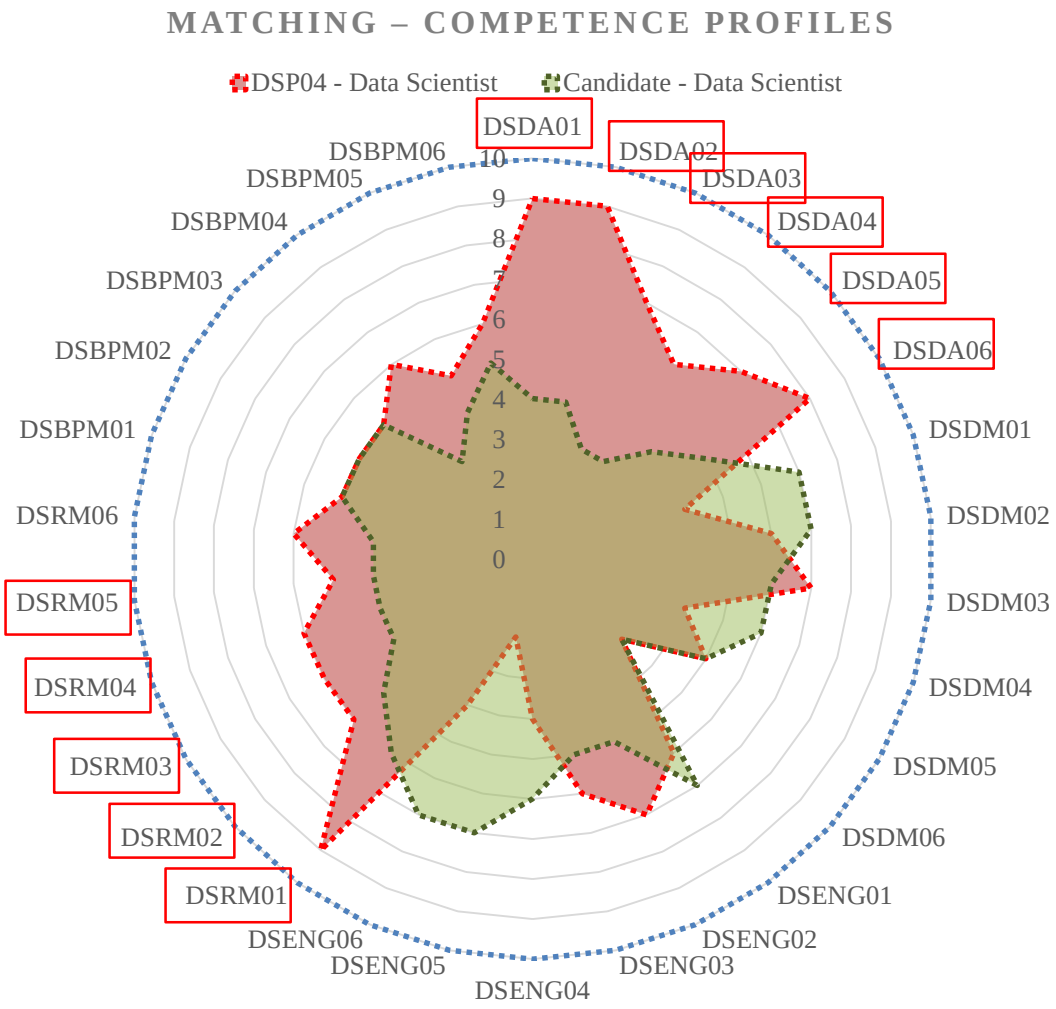
From Competences and DSP Profiles

to Learning Outcomes (LO) and

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

- EDSF allow for customized educational courses and training modules design

Individual Competences Benchmarking

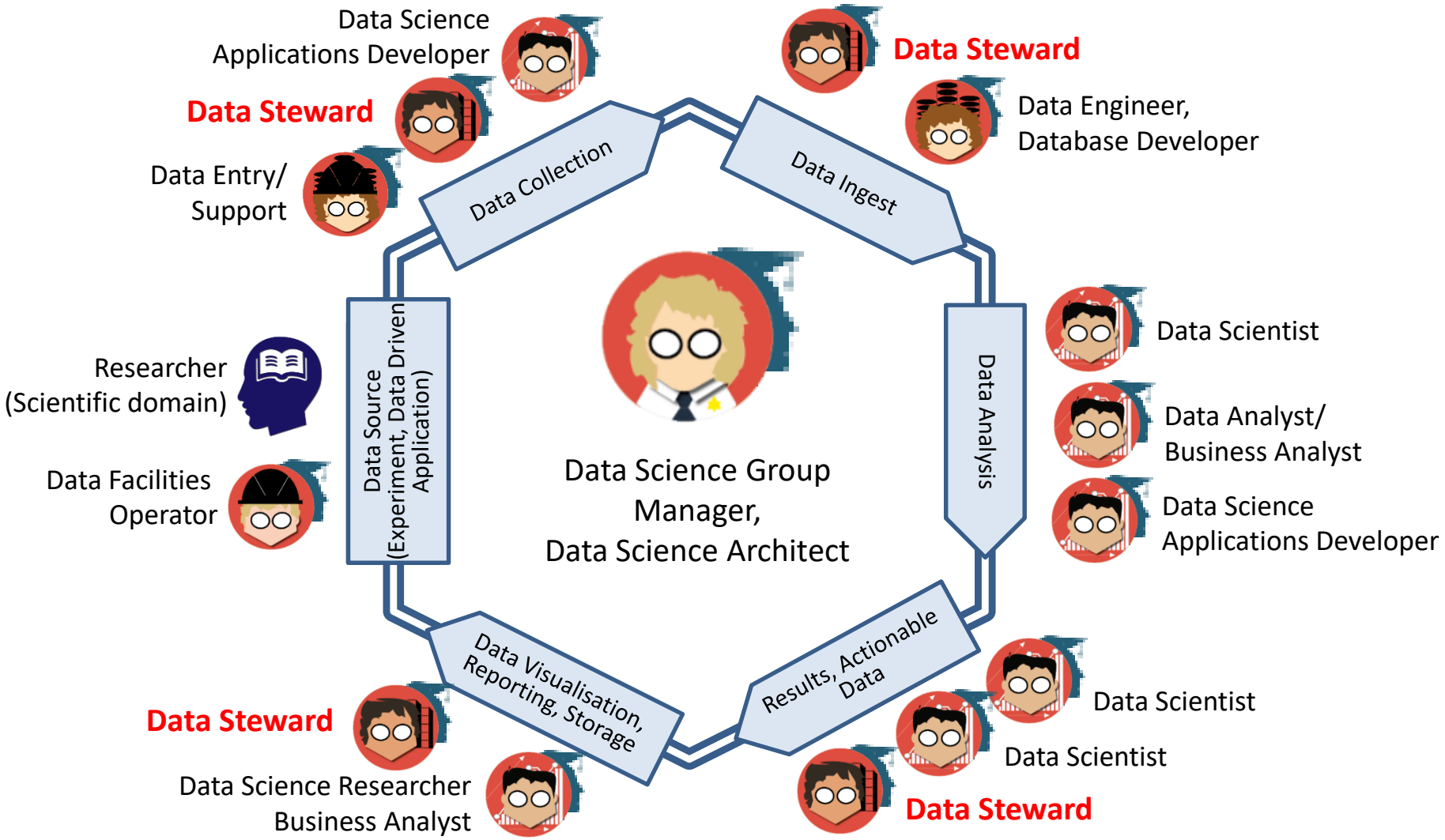


Individual Education/Training Path based on Competence benchmarking

- Red polygon indicates the chosen professional profile: Data Scientist (general)
- Green polygon indicates the candidate or practitioner competences/skills profile
- Insufficient competences (gaps) are highlighted in *red*
  - *DSDA01 – DSDA06 Data Science Analytics*
  - *DSRM01 – DSRM05 Data Science Research Methods*
- Can be use for team skills match marking and organisational skills management

[ref] For DSP Profiles definition and for enumerated competences refer to EDSF documents CF-DS and DSP Profiles.

Building a Data Science Team



Data 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

- (Managing) Data Science Architect (1)
- Data Scientist (1), Data Analyst (1)
- Data Science Application programmer (2)
- Data Infrastructure/facilities administrator/operator: storage, cloud, computing (1)
- **Data stewards**, curators, archivists (3-5)

Estimated: Group of 10-12 data specialists for research institution of 200-300 research staff.

Growing role and demand for Data Stewards and data stewardship

Data Stewards – A rising new role in Data Science ecosystem

- Data Stewards as a key bridging role between Data Scientists as (hard)core data experts and scientific domain researchers (HLEG EOSC report)
- Current definition of Data Steward (part of Data Science Professional profiles)
  - Data Steward is a data handling and management professional whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation.
  - Data Steward creates data model for domain specific data, support and advice domain scientists/ researchers during the whole research cycle and data management lifecycle.

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

Collaboration with the Research Data Alliance (RDA) on developing model curriculum on Research Data Literacy:
- Modular, Customisable, Localised, Open Access
- Supported by the network of trainers via resource swap board

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)

Data Science education and training in Europe: Roadmap recommendations

A. Policy recommendations for the EC and Member States

- R1. Critical skills management and training
- R2. Gender balance and multi-cultural environment
- R3. Common data literacy
- R4. Data driven technology and education divide.
- R5. European studies on demand and role of Data Science and Analytics skills
- R6. Include the above actions in the European Digital Single Market Scoreboard

B. Recommendations to universities and professional training organisations

- R7. EDSF adoption by universities
- R8. Addressing Data Science professional and workplace skills in university curricula
- R9. Multiple delivery form
- R10. Sharing experience, courses and instructors
- R11. Supporting technical infrastructure for Data Science and data related education and professional training

C. Recommendations for Research (e) Infrastructures

- R12. Critical skills management in Research (e) Infrastructures

D. Required support and contribution to the standardisation bodies

- R13. The following are essential measures to achieve EDSF support existing and new standards.

Next steps (1) – Further development and exploitation

- EDSF Release 3 (October 2017)
  - Fully enumerated CF, BoK, MC, DSPP in machine readable format
- Development of EDSF based tools:
  - Self assessment and Market monitoring
- Certification Framework for at least two levels of Data Science competences proficiency: Associates and Professionals
  - Data Science knowledge and competences for decision makers

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

Next steps (2) – Community Activities and Initiatives

- Data Science Manifesto – Primarily focused on professional and ethical issues in Data Science, new type of professional
  - Professional and ethical issues as a primary focus

- Inter-universities initiative “Data Science for UN’s Sustainable Development Goals” to focus in-curricula research (projects) on UN priority goals

- Build a wider network of early adopters, champions and ambassadors in Europe and world wide
  - EDISON team information and advice support will continue through out 2017-2018

Questions and discussion

Links to EDISON Resources

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

Other related links

- Amsterdam School of Data Science
  - https://www.schoolofdatascience.amsterdam/
  - https://www.schoolofdatascience.amsterdam/education/
- Research Data Alliance interest Group on Education and Training on Handling of Research Data (IG-ETHRD)
  - https://www.rd-alliance.org/groups/education-and-training-handling-research-data.html
- Final Report on European Data Market Study by IDC (Feb 2017)
  - https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy
- PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017)
  - http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent
- Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017)
  - http://www.bhef.com/publications/quant-crunch-how-demand-data-science-skills-disrupting-job-market
  - https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF
- Millennials at work: Reshaping the workspace (2016)
  - https://www.pwc.com/m1/en/services/consulting/documents/millennials-at-work.pdf

Additional materials

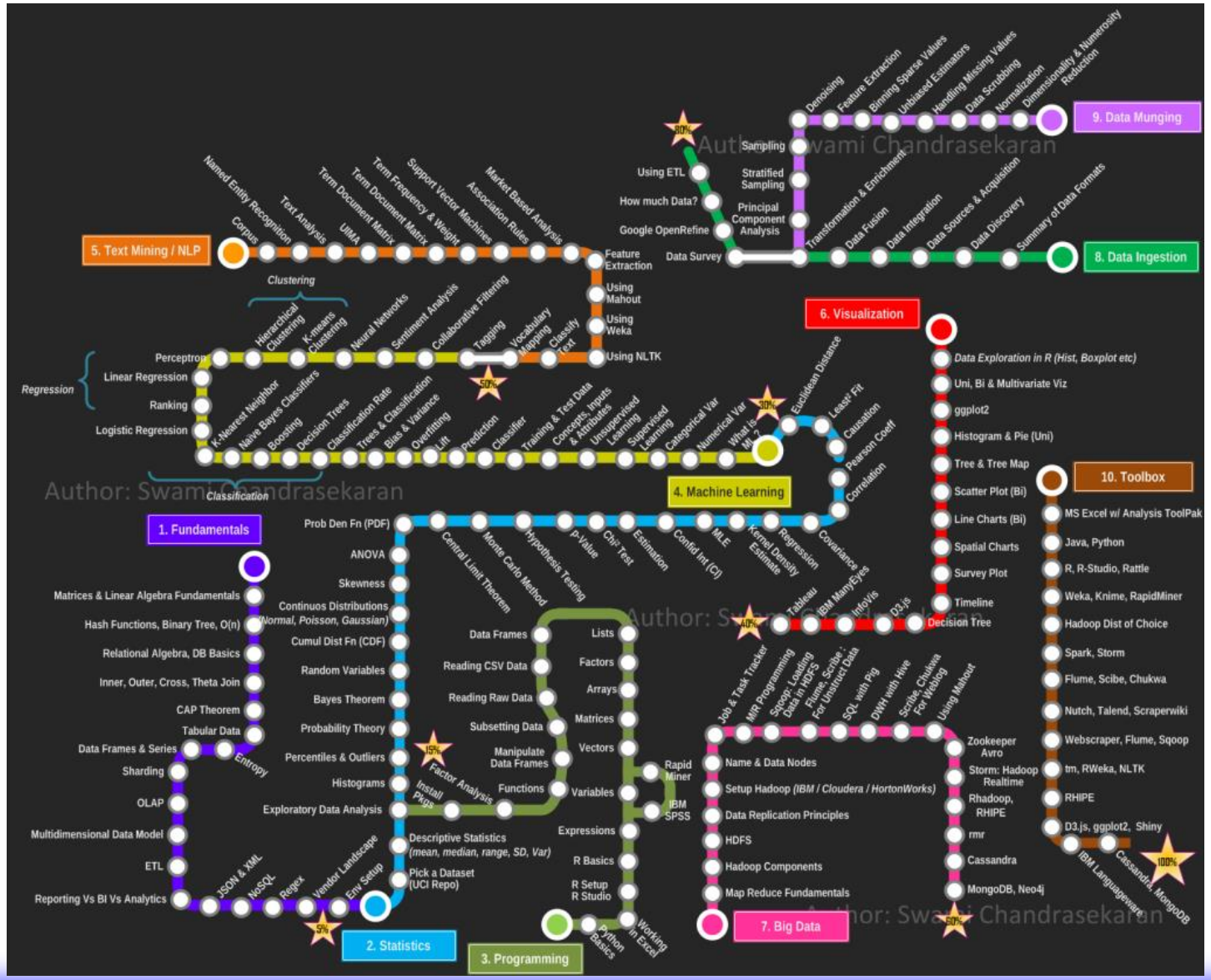
EDSF Recognition, Endorsement and Implementation

- **DARE (Data Analytics Rising Employment)** project by APEC (Asia Pacific Economic Cooperation)
  - DARE project Advisory Council meeting 4-5 May 2017, Singapore
- **PcW and BHEF Report "Investing in America's data science and analytics talent"** April 2017
  - Quotes EDSF and Amsterdam School of Data Science
- **Dutch Ministry of Education recommended EDSF** as a basis for university curricula on Data Science
  - Workshop "Be Prepared for Big Data in the Cloud: Dutch Initiatives for personalized medicine and health research & toward a national action programme for data science training", Amsterdam 28 June 2016
  - Currently working with Dutch Gov on re-skilling IT/data workers for DSA competences
- **European Champion Universities network**
  - 1st Conference (13-14 July, UK)
  - 2nd Conference (14-15 March, Madrid, Spain)
  - 3rd Conference 19-20 June 2017, Warsaw

How to become a Data Scientist

- Understand required Data Science and Analytics competences and skills
- Build your own learning path
  - Assess your knowledge and start from basics
  - Statistics is foundation of Data (Science) Analytics
    - Develop statistical/probabilistic thinking
    - Difference between Data Science and statistics
  - Learn from others experience: read blogs, join forums and communities
  - Decide about academic degree, professional certificate, self-education/training, join local Meetup
- Start applying for job
  - Remember variety of Data Scientist roles and profiles

Becoming a Data Scientist by Swami Chandrasekaran (2013) http://nirvacana.com/thoughts/becoming-a-data-scientist/



- Good and practical advice how to learn Data Science, step by step
- Follow the route