EDISON Data Science Framework (EDSF)

Data Science Professional Education and Training

EDISON Project value proposition and legacy

Yuri Demchenko, EDISON Project
University of Amsterdam

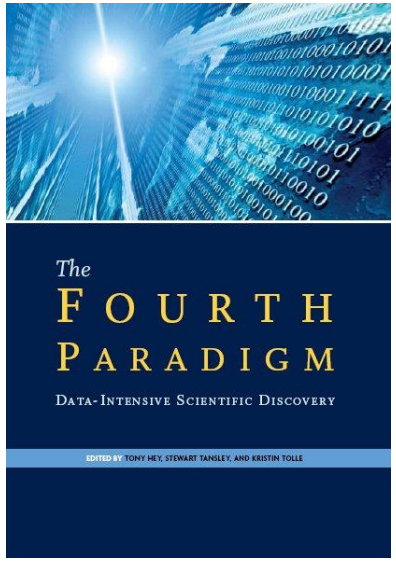
EDISON Initiative, February 2025

EDISON Project (2015-2017)
Grant 675419 (INFRASUPP-4-2015: CSA)

Outline

- Background: Data driven research and demand for new skills
  - Foundation, 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
  - Managing Data Science Teams
- Data Science Body of Knowledge and Model Curriculum
- Use of EDSF and Example curricula
  - Competences assessment
  - Building Data Science team
- Roadmap recommendations
- References and additional materials

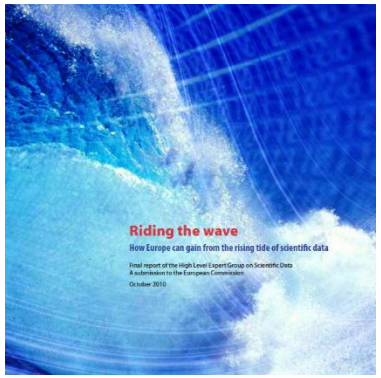
Visionaries and Drivers (2015): Seminal works, High level reports, Activities



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)

https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

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** [ref]

- **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.

[ref] https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

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-TRAIN: Training of data stewards capable of providing FAIR data services
  - FAIRdICT: Top Sector Health collaboration with top team ICT
- 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

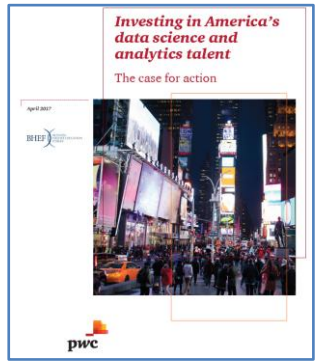
https://www.dtls.nl/fair-data/
https://www.dtls.nl/fair-data/go-fair/
https://www.dtls.nl/fair-data/fair-data-training/

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

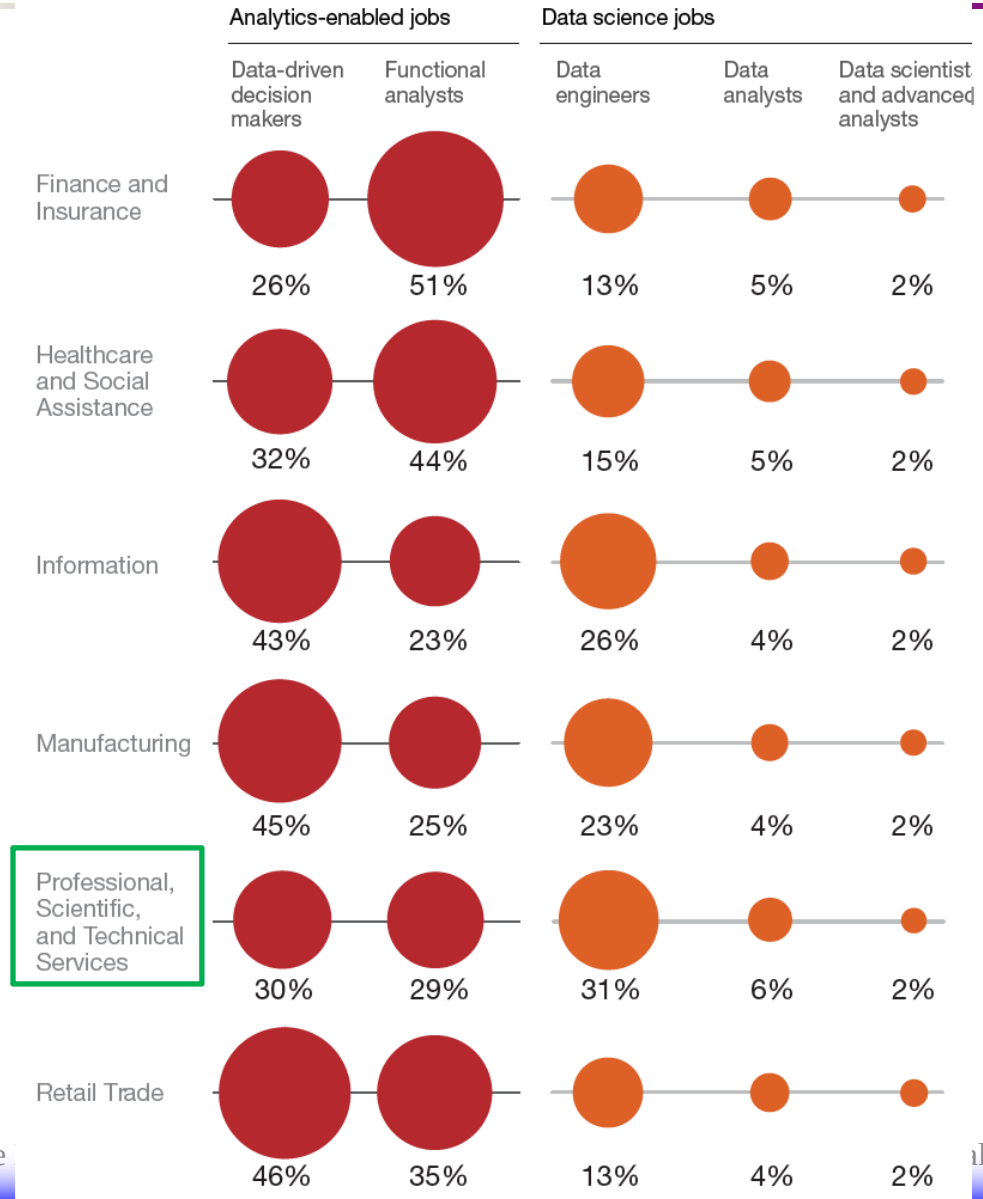
- Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017) - Edited
  - https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF
  - DSA enabled jobs takes 45-58 days to fill: 5 days longer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



Citing EDISON and EDSF

Influenced by EDISON

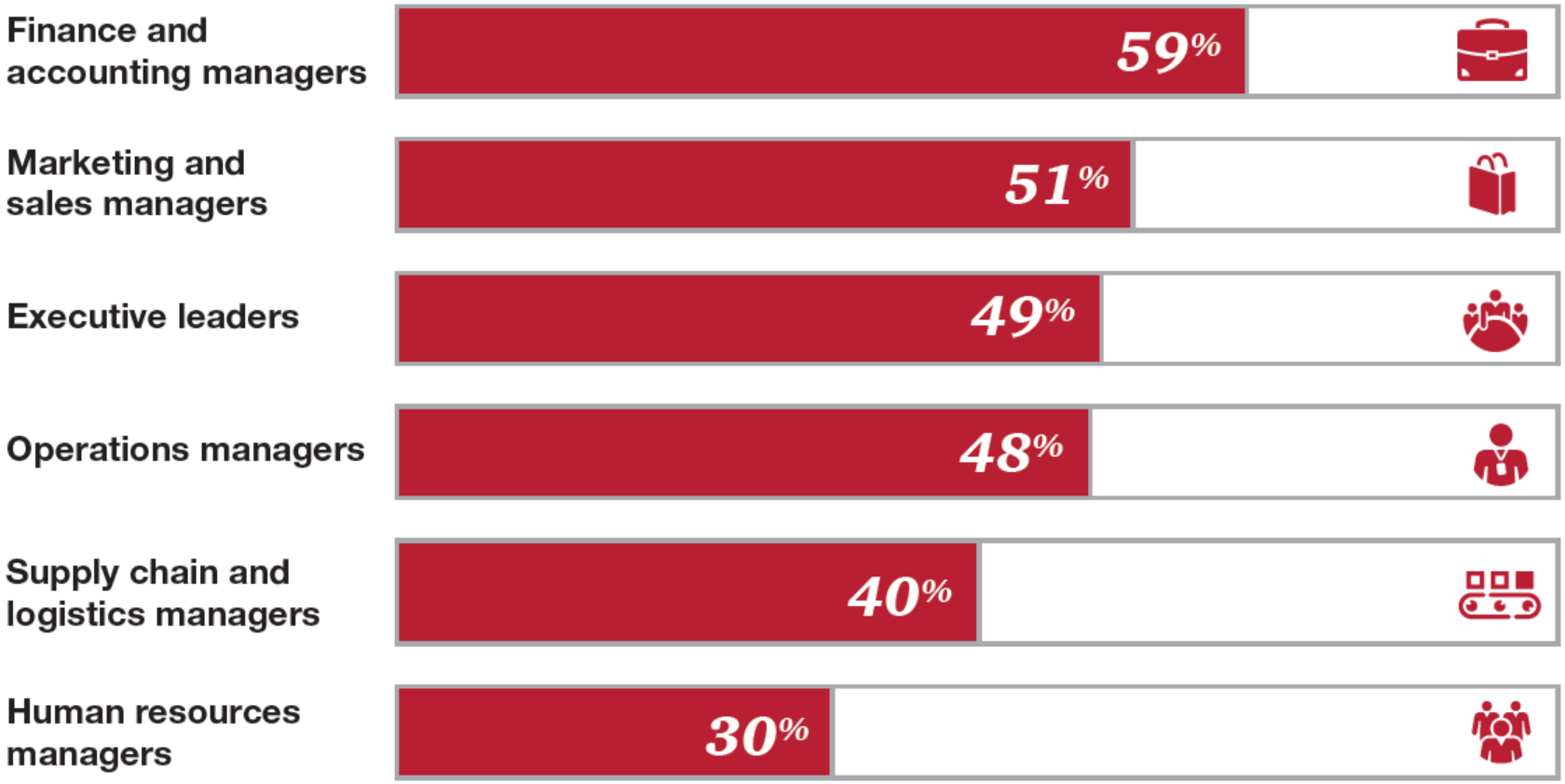
PwC&BHEF: Demand for DSA enabled jobs



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

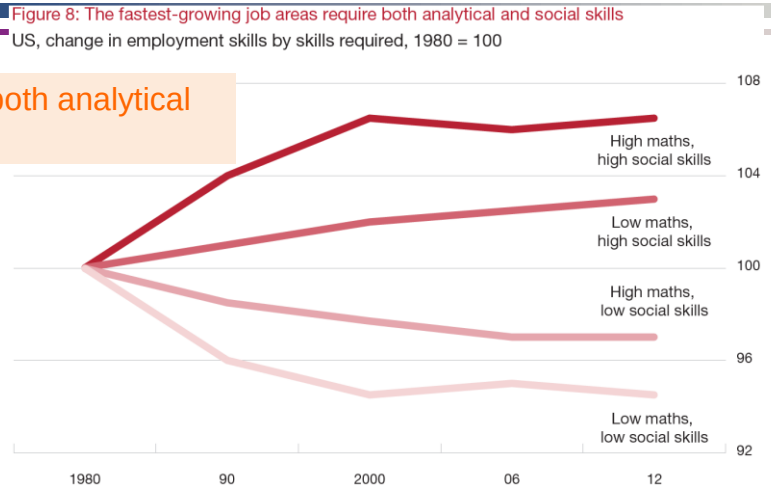


Percent of employers who say data science and analytics skills will be 'required of all managers' by 2020

- 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

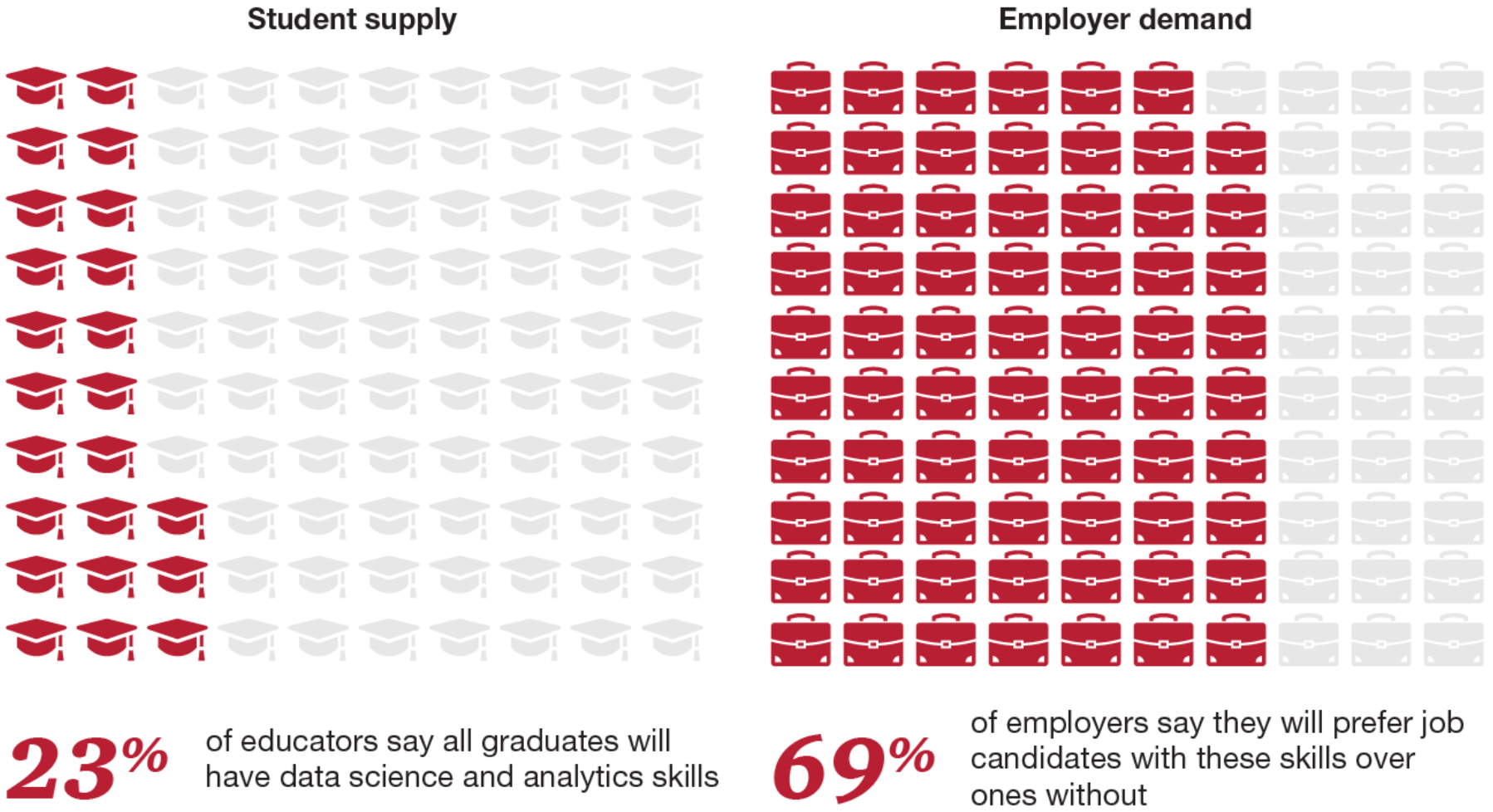
78% Cognitive flexibility

74% Cross-disciplinary ability

Source: Business Roundtable (2017).

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

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

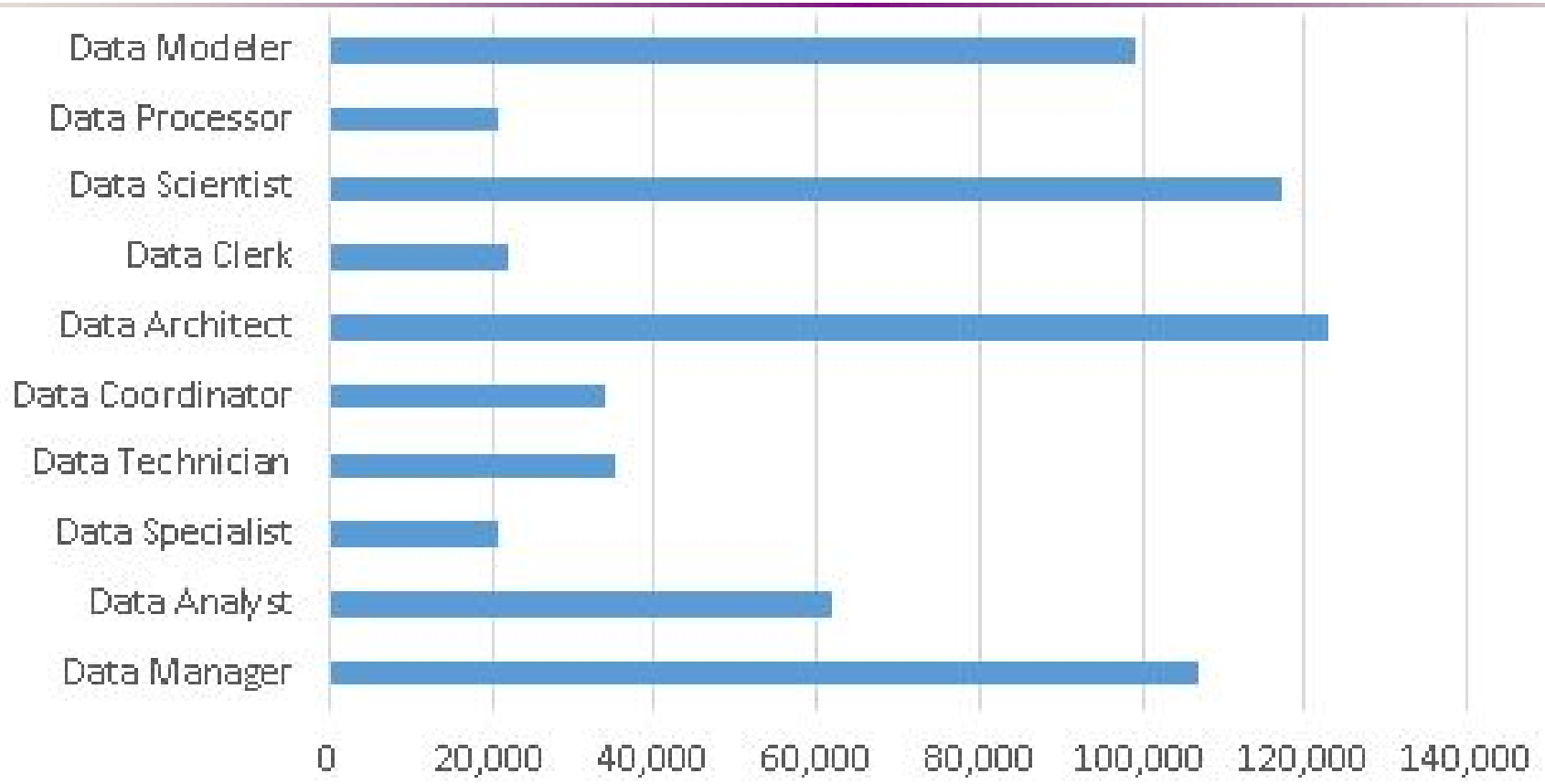


IBM&BGT: DSA Jobs Time to Fill and Salary (2016-2017)

DSA Framework Category	Top Industries (by Demand Volume)	Average Time to Fill (Days)	Average Annual Salary
Data-Driven Decision Makers	Professional Services	50	$96,845
	Finance & Insurance	37	$98,131
	Manufacturing	43	$93,641
Functional Analysts	Finance & Insurance	35	$71,937
	Professional Services	48	$69,135
	Manufacturing	39	$72,571
Data Systems Developers	Professional Services	51	$82,447
	Finance & Insurance	35	$87,039
	Manufacturing	43	$81,138
Data Analysts	Professional Services	47	$74,917
	Finance & Insurance	31	$83,209
	Manufacturing	41	$72,742
Data Scientists & Advanced Analysts	Professional Services	51	$97,457
	Finance & Insurance	43	$106,610
	Manufacturing	45	$92,543
Analytics Managers	Finance & Insurance	38	$113,754
	Professional Services	53	$107,185
	Manufacturing	40	$106,926

- On average, DSA jobs in Professional Services remain open for 53 days, eight days longer than the overall DSA average. (IBM, BGT 2017 Study)

Closer look at Data related Jobs and Salaries (2016)



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

OECD and UN on Digital Economy and Data Literacy

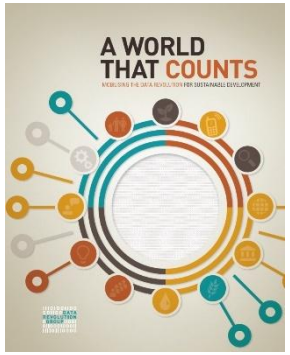
OECD (Organisation for Economic Cooperation and Development)

- 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
http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/ICCP/IIS(2015)10/FINAL&docLanguage=En

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 and Industry 4.0
- **Data literacy** = critically analyse data collected and data visualised



PwC study: Millennials at work (2016) – 1

https://www.pwc.com/co/es/publicaciones/assets/millennials-at-work.pdf

Confirmed results of previous studies:

- Loyalty-lite to company
  - The power of employer brands and the waning importance of corporate responsibility
- A time of compromise: benefit from individual package negotiation
- Development and work/life balance are more important than position or salary
  - Work/life balance and diversity promises are not being kept
- Financial reward is secondary but cash bonuses are valued
- A techno generation avoiding face time and prefer network communication
- Moving up the ladder faster expectation but often not confirmed by hard work required
- Generational communication but not without tensions

PwC study: Millennials at work (2016) – 2

https://www.pwc.com/co/es/publicaciones/assets/millennials-at-work.pdf

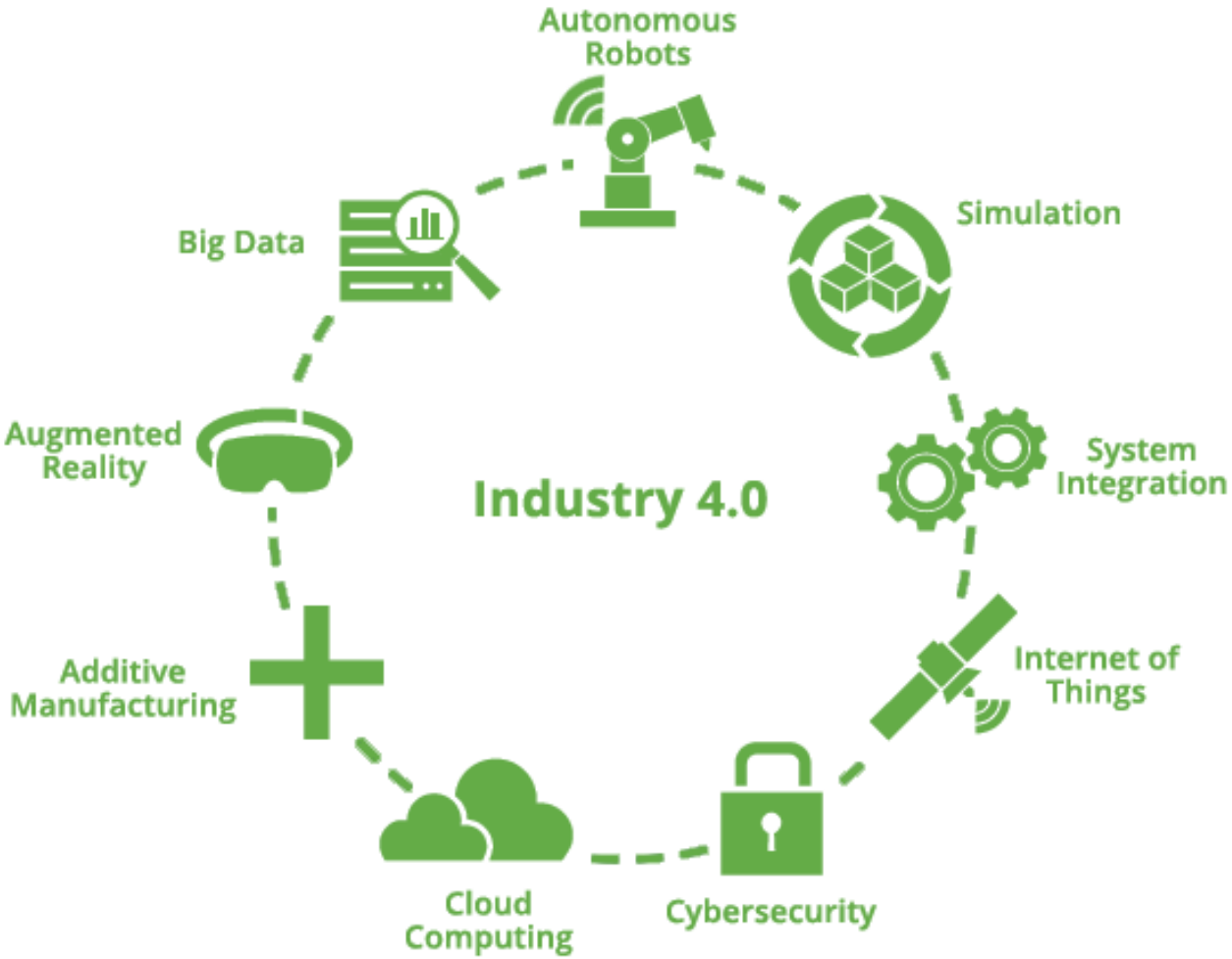
- What organisation is an attractive employer?
  - Opportunities for career progression
  - Competitive wages/other financial incentives
  - Excellent training/development programmes
- Factors most influenced decision to accept your current job?
  - The opportunity for personal development
  - The reputation of the organisation
  - The role itself
- Which three benefits would you most value from an employer?
  - Training and development
  - Flexible working hours
  - Cash bonuses

What can employers do?

Business leaders and HR need to work together to:

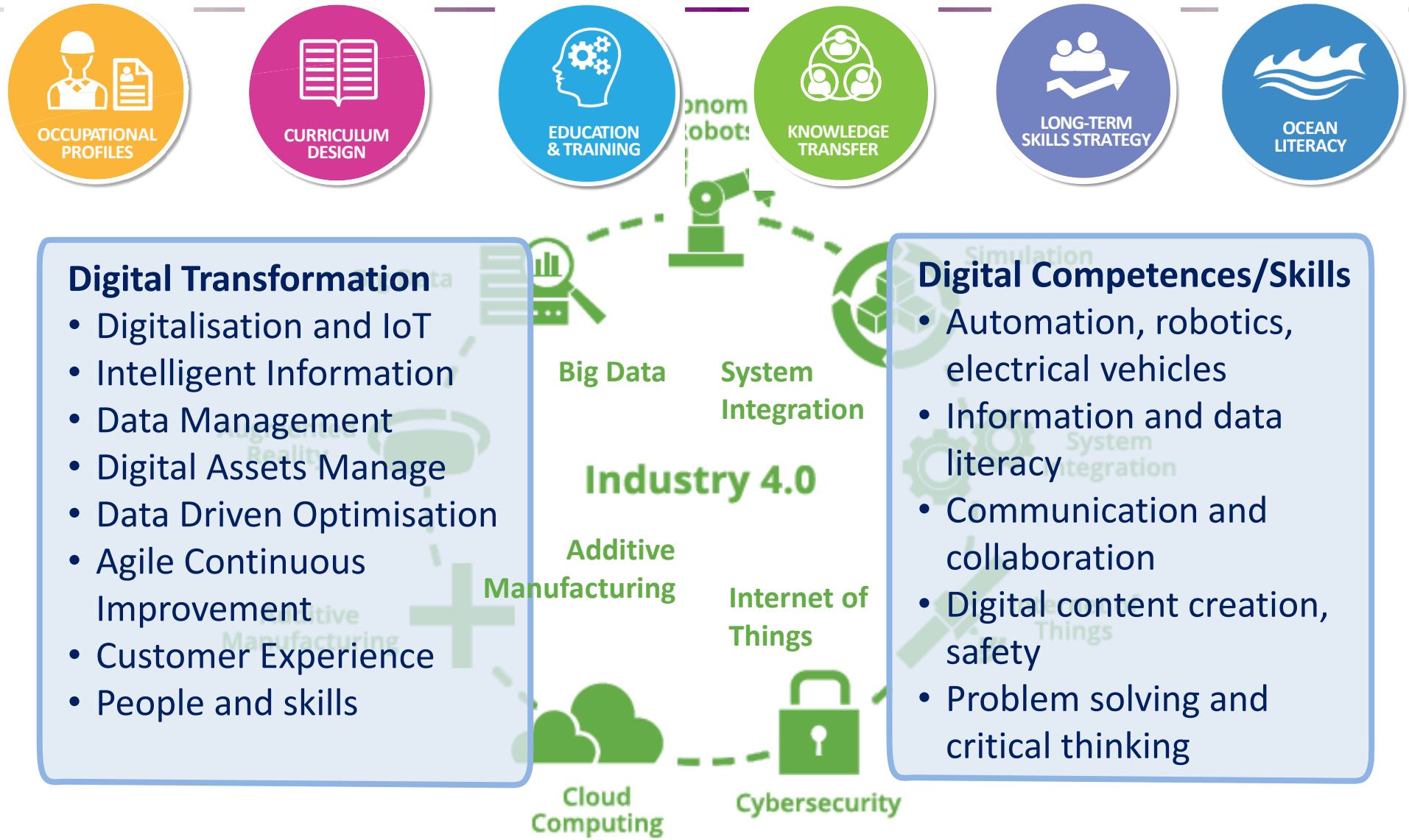
- Understand this generation
- Get the ‘deal’ right
- Help millennials grow
- Feedback, feedback and more feedback
- Set them free
- Encourage learning
- Allow faster advancement
- Expect millennials to go

Industry 4.0 and demand for new skills



The Fourth Industrial Revolution, which includes developments in previously disjointed fields such as artificial intelligence and machine-learning, robotics, nanotechnology, 3-D printing, and genetics and biotechnology, will cause widespread disruption not only to business models but also to labour markets over the next five years, with enormous change predicted in the skill sets needed to thrive in the new landscape. This is the finding of a new report, *The Future of Jobs*, published today by the World Economic Forum.

Maritime Industry Digital Transformation and Skills Strategy – Toward Industry 4.0



Digital Transformation

- Digitalisation and IoT
- Intelligent Information
- Data Management
- Digital Assets Manage
- Data Driven Optimisation
- Agile Continuous Improvement
- Customer Experience
- People and skills

Digital Competences/Skills

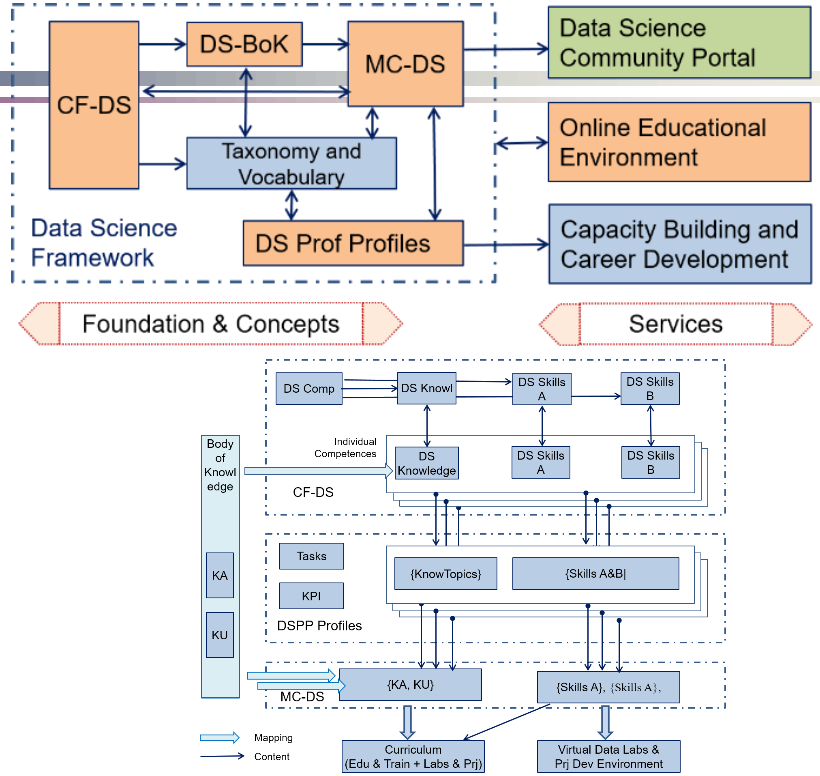
- Automation, robotics, electrical vehicles
- Information and data literacy
- Communication and collaboration
- Digital content creation, safety
- Problem solving and critical thinking

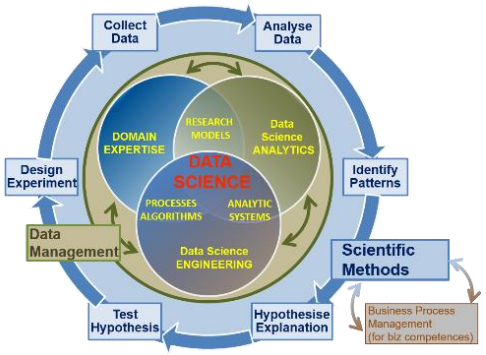
Challenge for Education: Sustainable ICT and Data Skills Development

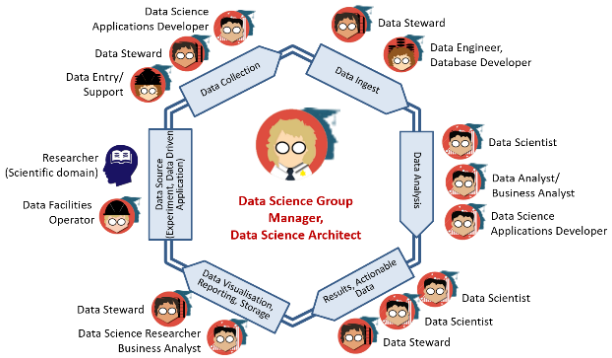
- Educate vs Train
  - Training is a short term solution
  - Education is a basis for sustainable skills development
- Technology focus changes every 3-4 years
  - Study: 50% of academic curricula are outdated at the time of graduation
- Lack of necessary skills leads to *underperforming projects* and 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

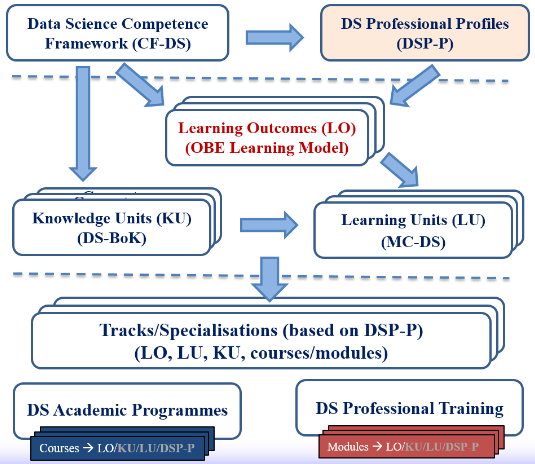
EDISON Products for Data Science Skills Management and Curriculum Design

- EDISON Data Science Framework (EDSF)
  - Release 3 components CF-DS Competence Framework, DS-BoK Body of Knowledge, MC-DS Model Curriculum, DSPP Professional Profiles
  - Compliant with EU standards on competences and professional occupations e-CFv4.0, ESCO
- Skills development and career management for Core Data Experts and related data handling professions
- Academic programmes and professional training courses (self) assessment and design
- Individual competences benchmarking and Data Science team design
- Cooperation with International professional organisations IEEE, ACM, BHEF, APEC (AP Economic Cooperation)

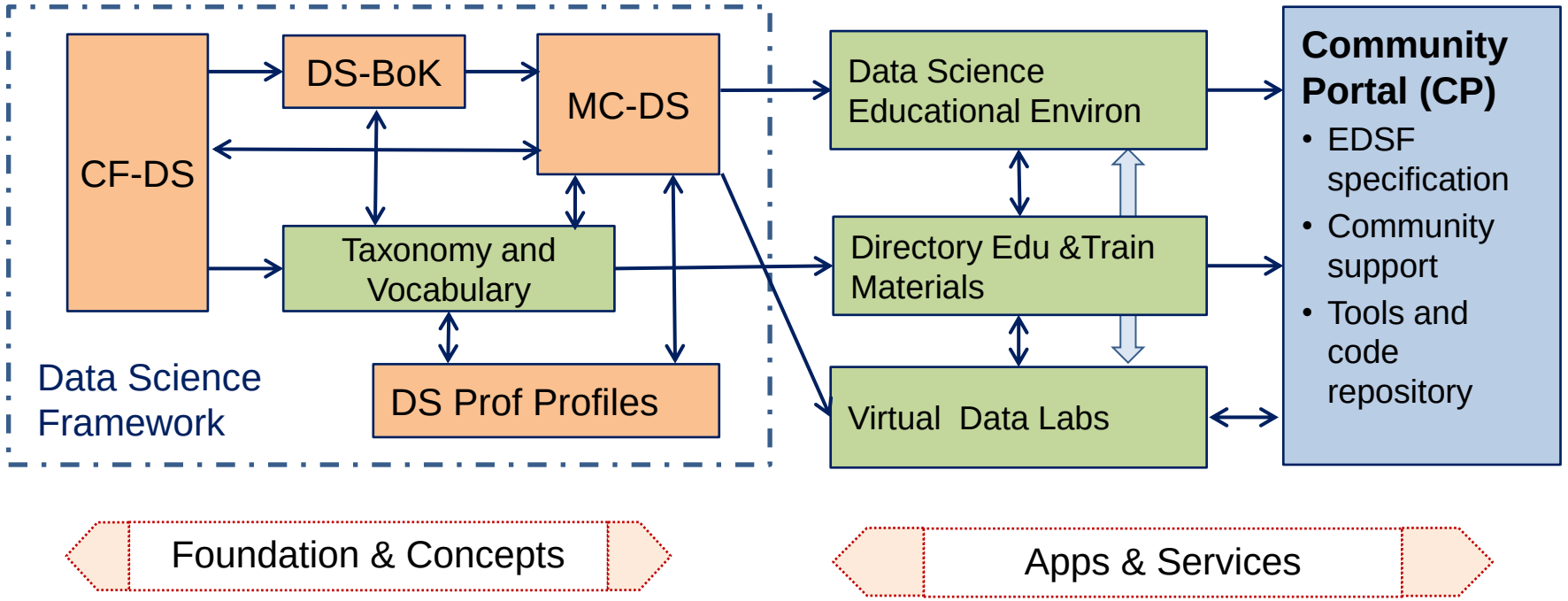








EDISON Data Science Framework (EDSF release 4) – Core components and community maintained services



EDISON Framework core components and documents

- CF-DS – Data Science Competence Framework (Part 1)
- DS-BoK – Data Science Body of Knowledge (Part 2)
- MC-DS – Data Science Model Curriculum (Part 3)
- DSPP – Data Science Professional profiles (Part 4)
- Data Science Taxonomies and Scientific Disciplines Classification

Applications and Services

- Virtual Data Science Labs
- Data Science Educational Environment
- Directory of edu & train resources
- Community Portal – currently github

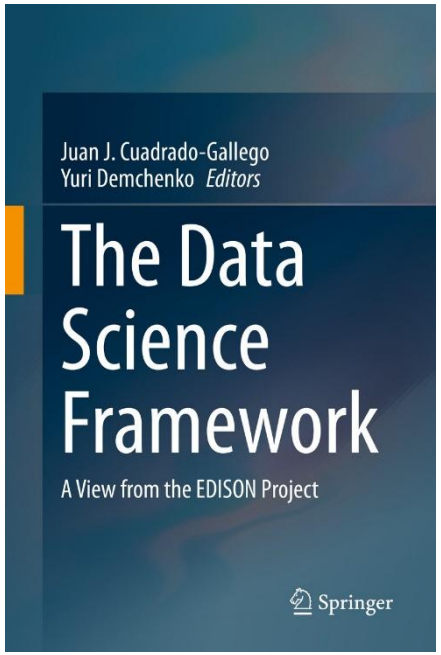
EDSF Release 4.0 – Published December 2022

- Call for Use cases contribution – Deadline 30 Sept 2022

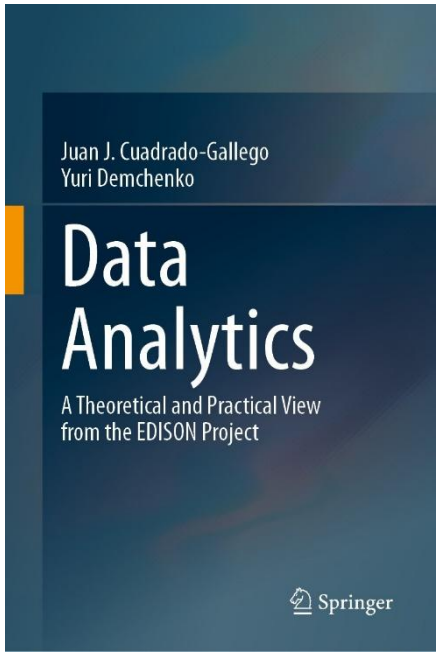
- EDSFr4 Design workshop – 12-22 Nov 2019, Amsterdam
  - Next to FAIRsFAIR Thematic group meeting

- New developments
  - **New Part 5: Use cases and Guidelines**
    - Including definition of 21st Century Skills and Data Science professional skills
    - Including Data Stewardship competences definition (by EOSCpilot and FAIRsFAIR projects)
    - Including summary and links to known curricula and resources
- Further Developments
  - Using EDSF methodology for analysing job market for trending AI/GenAI and Green Competences
  - Moving EDSF to ontology definition format
  - New Part 6: Vocabulary and terms definition -

Books Inspired by EDSF



2020



2023

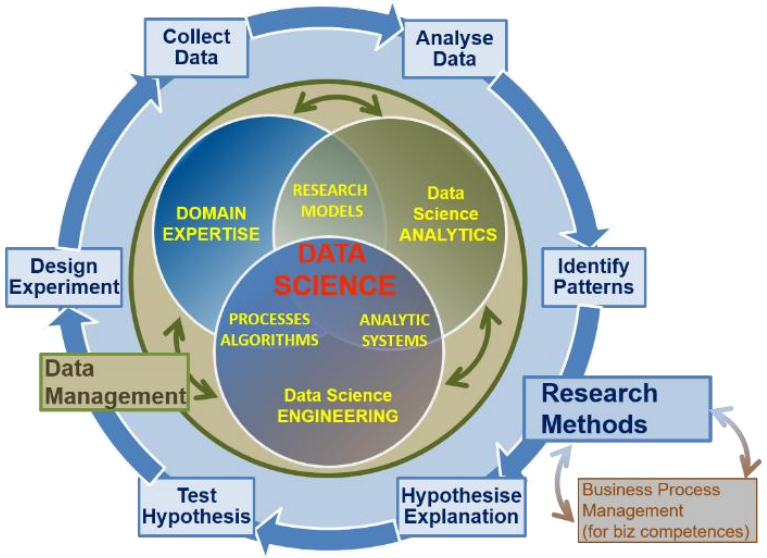
October 2024

- The Data Science Framework: A View from the EDISON Project, Editors: Juan J. Cuadrado-Gallego, Yuri Demchenko, Springer 2020. DOI: https://doi.org/10.1007/978-3-030-51023-7
- Data Analytics: A Theoretical and Practical View from the EDISON Project. Authors: Juan J. Cuadrado-Gallego, Yuri Demchenko, Springer 2023. DOI: https://doi.org/10.1007/978-3-031-39129-3
- Big Data Infrastructure Technologies for Data Analytics: A Theoretical and Practical View from the EDISON Project. Authors: Yuri Demchenko, Juan J. Cuadrado-Gallego, Oleg Chertov, Marharyta Aleksandrova, Springer 2024

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.*



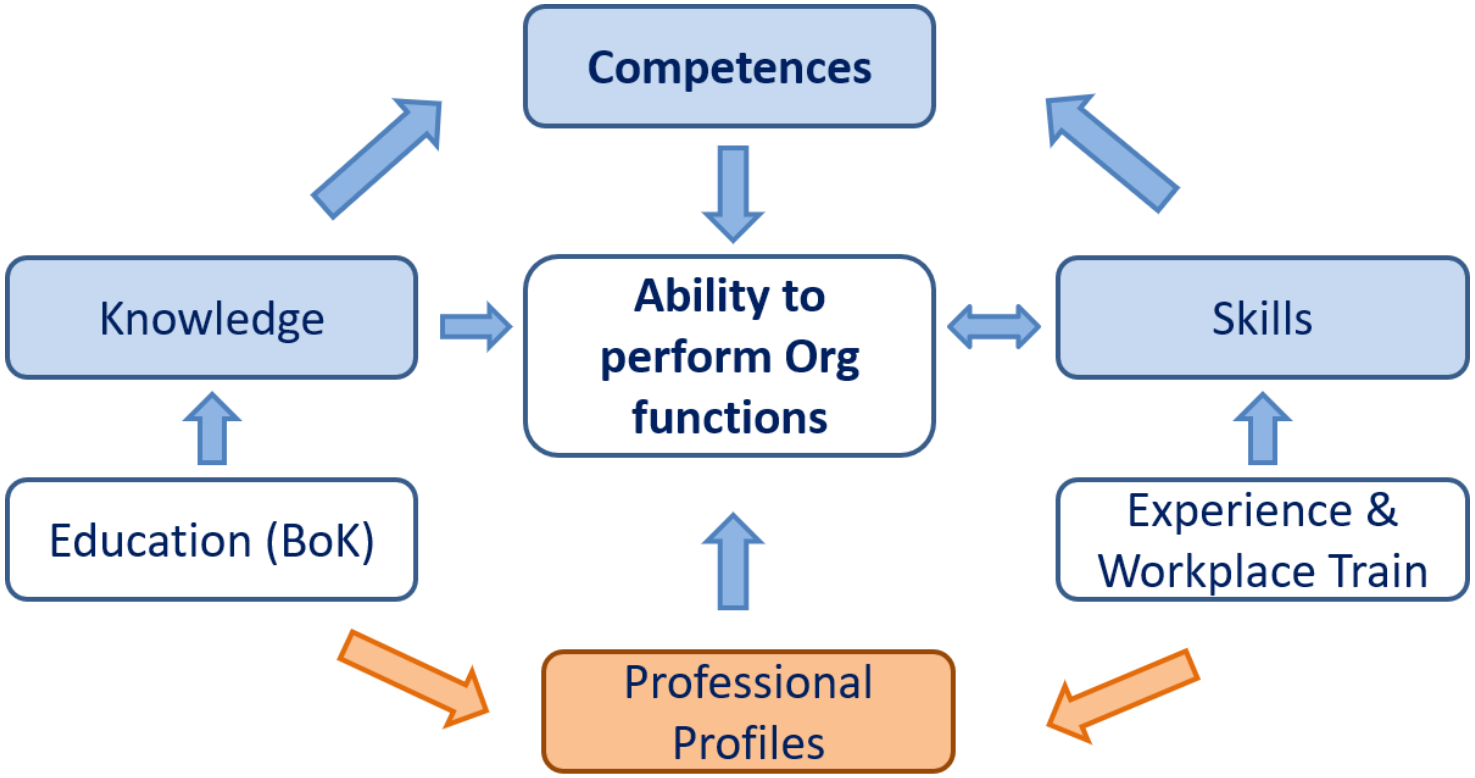
- Core Data Science competences and skills groups
  - **Data Science Analytics** (including Statistical Analysis, Machine Learning, Business Analytics)
  - **Data Science Engineering** (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
  - **Domain 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

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

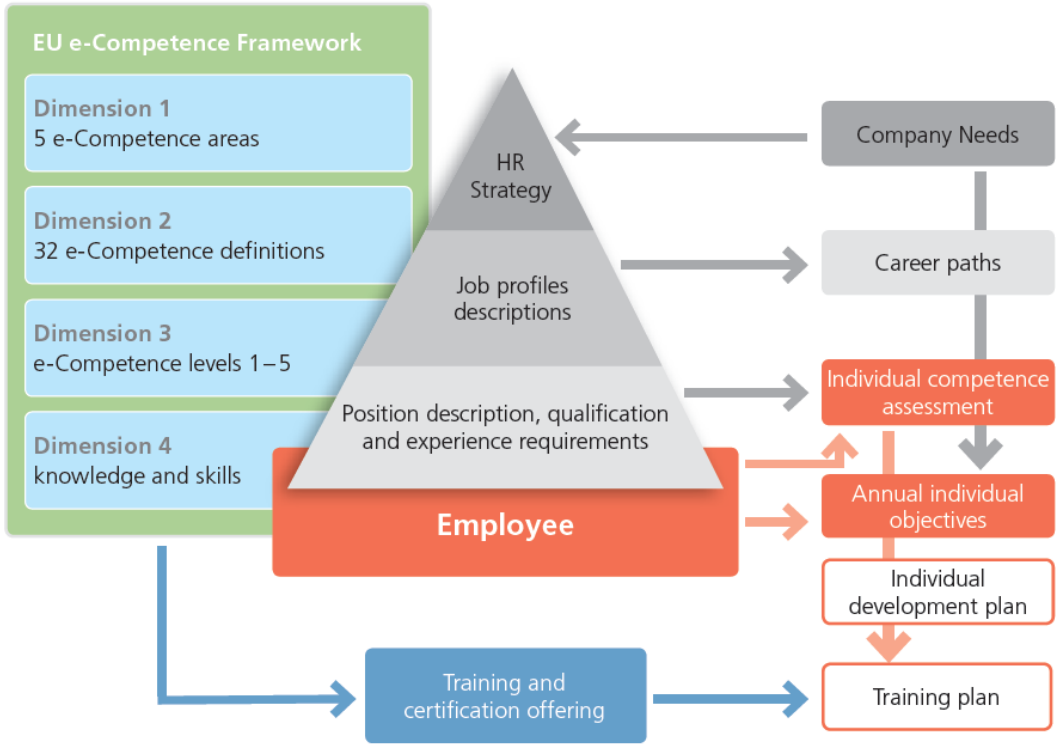
Competences Map to Knowledge and Skills

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



CF-DS/EDSF Approach: Compatibility with e-CFv3.0 structure and 4-dimensional model

- Competence Framework for Data Science (CF-DS) definition will be built based on European e-Competence framework for IT (e-CFv3.0)
  - Linking scientific research lifecycle, organizational roles, competences, skills and knowledge
  - Defining Data Science Body of Knowledge (DS-BoK)
  - Mapping CF-DS and DS-BoK to academic disciplines



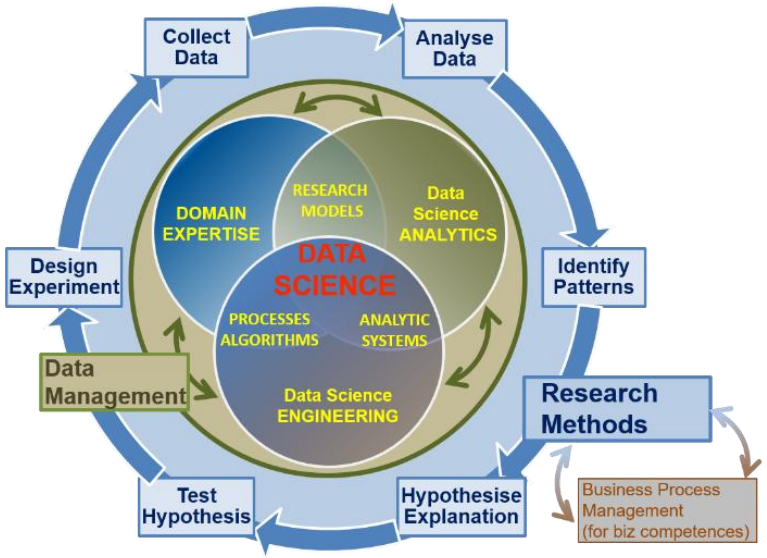
- Multiple use of e-CFv3.0 within ICT organisations
- Provides basis for individual career path, competence assessment, training and certification

- 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

Data Scientist definition

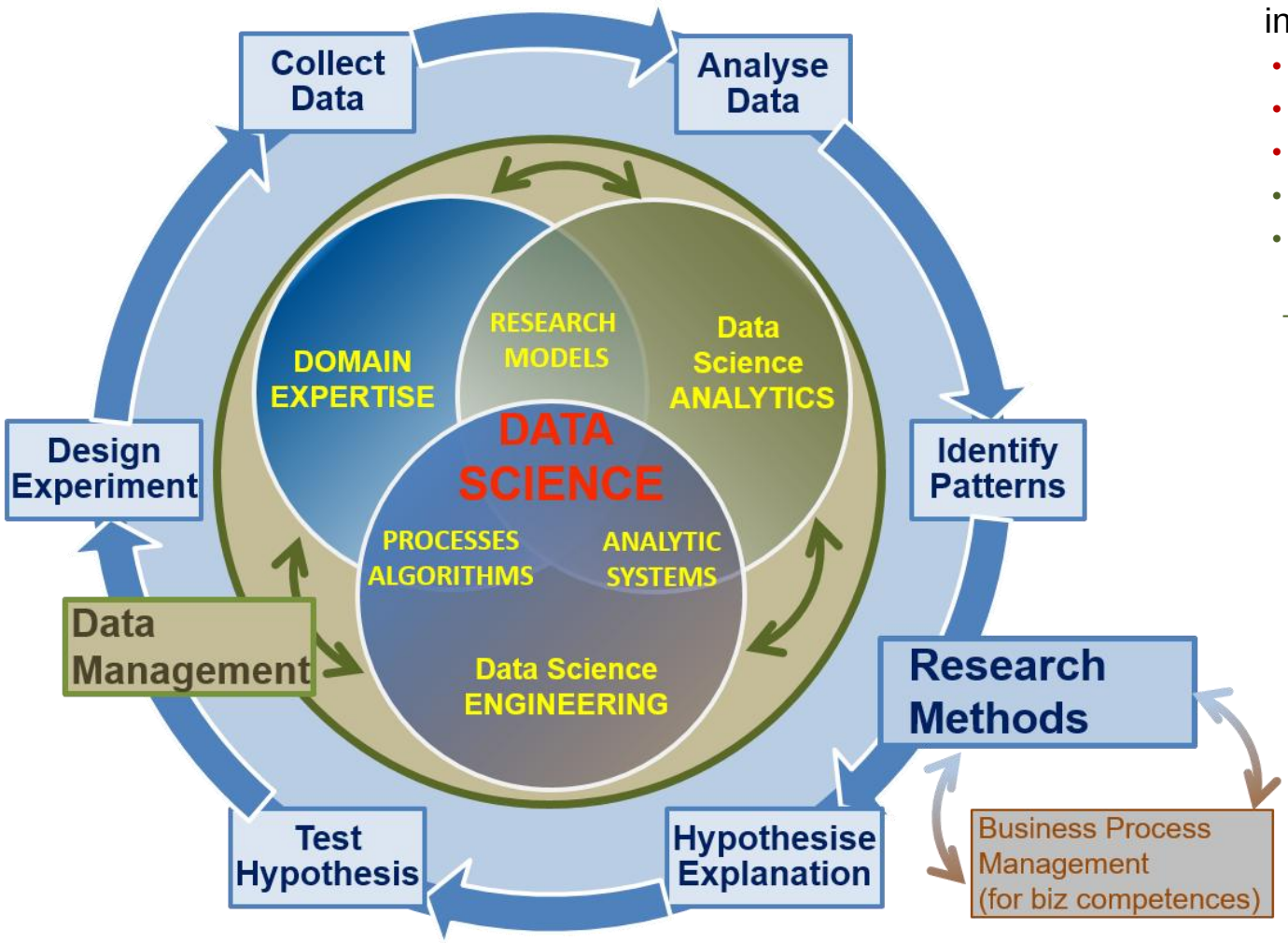
Based on the definitions by NIST SP1500 – 2015, extended by EDISON

- *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
- Research Methods and Project Management
  - Business Process Management (biz)

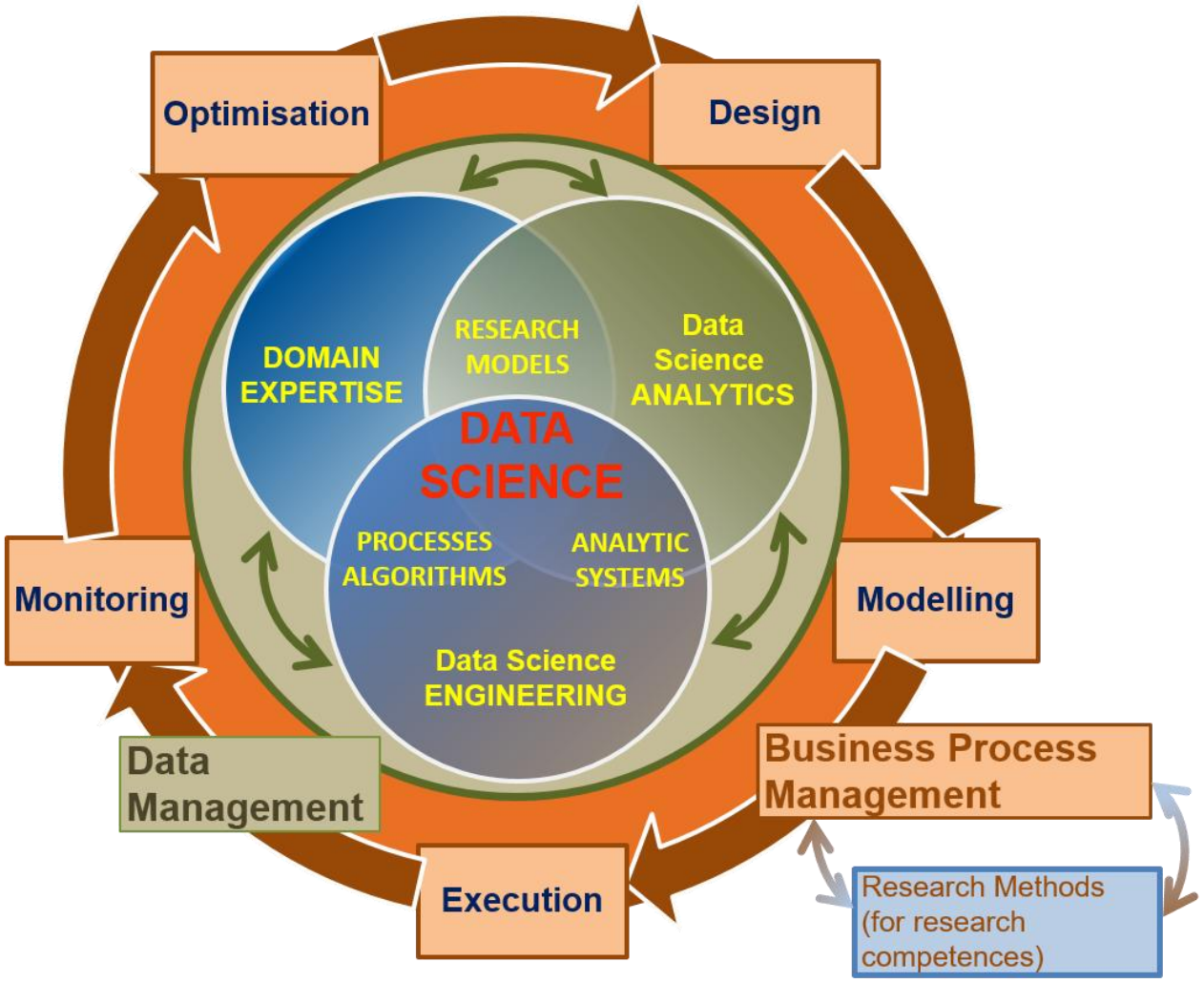
Scientific Methods

- 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
- Research Methods and Project Management
  - Business Process Management (biz)

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

Skills Type A – Based on knowledge acquired

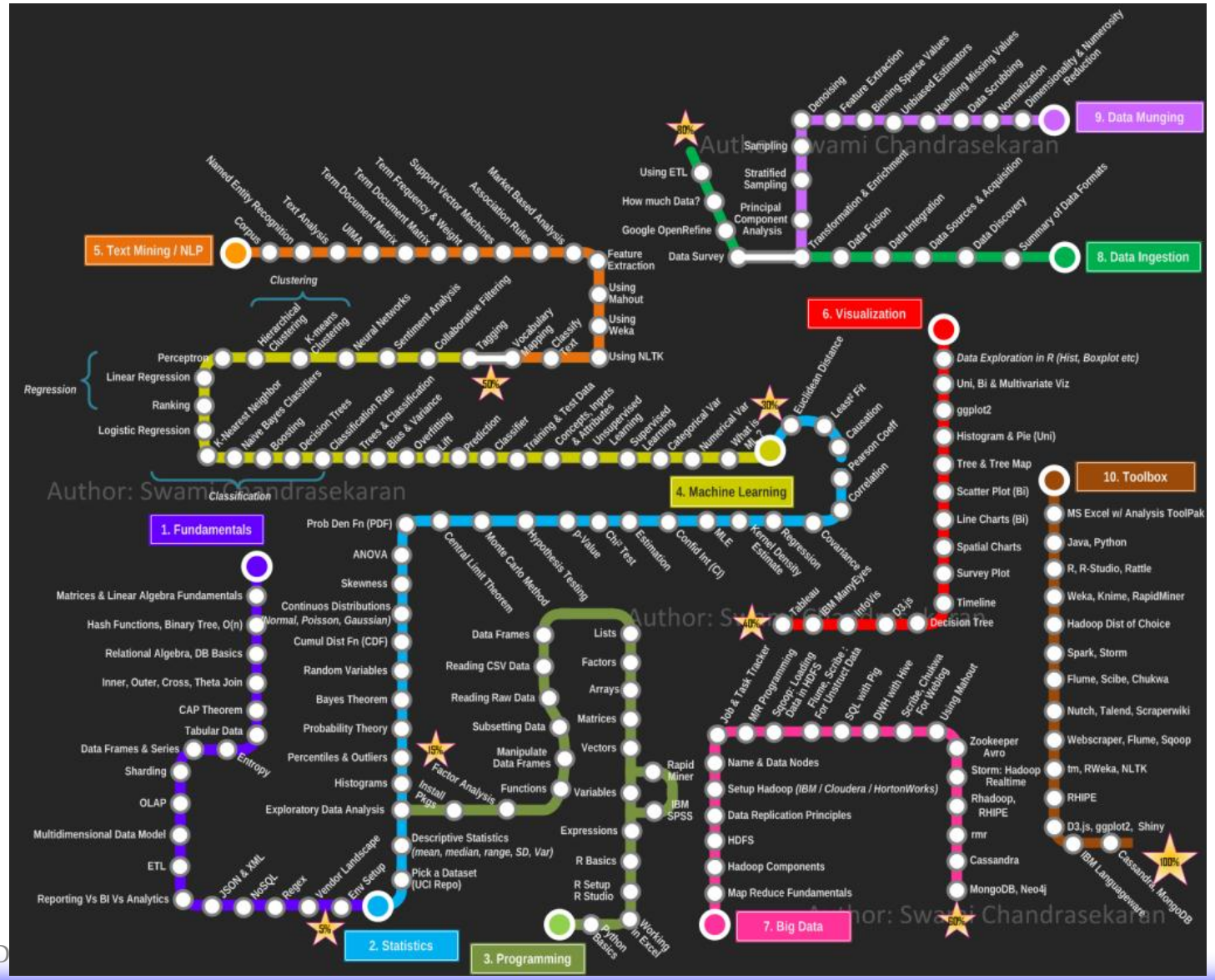
- **Group 1: Skills/experience related to competences**
  - Data Analytics and Machine Learning
  - Data Management/Curation (including both general data management and scientific data management)
  - Data Science Engineering (hardware and software) skills
  - Scientific/Research Methods or Business Process Management
  - Application/subject domain related (research or business)
- **Group 2: Mathematics and statistics**
  - Mathematics and Statistics and others

Skills Type B – Base on practical or workplace experience

- **Group 3: 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 4: Data analytics programming languages and IDE**
  - General and specialized development platforms for data analysis and statistics
- **Group 5: Soft skills and Workplace skills**
  - Data Science professional skills: Thinking and Acting like Data Scientist
  - 21st Century Skills: Personal, inter-personal communication, team work, professional network

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

Example Data Science Competences Definition Compliant with e-CFv3.0

Dimension 1 Competence Group	DSDA	Data Science Analytics
Dimension 2 Competence	DSDA04	Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval
Dimension 3 Proficiency level	Level 1 (Entry/Associate)	Level 1 (Professional)	Level 1 (Expert)
	Be familiar and be able to use different performance and accuracy metrics as part of used data analytics platforms	Select appropriate performance metrics and apply them for specific analytics applications. Develop new metrics and use it for fine tuning the used analytics solutions.	Not specifically defined. Advanced knowledge and experience.
Dimension 4 Knowledge	Knowledge ID	Knowledge unit definition
	KDSDA01	Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others
	KDSDA02	Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)
	KDSDA06	Predictive Analytics
	KDSDA11	Performance and accuracy metrics
	KDSDA14	Optimisation
	Skill ID	Skills definition
Skills Data Analytics methods and algorithms	SDSDA01	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)
	SDSDA04	Apply Predictive Analytics methods
	SDSDA09	Be able to use performance and accuracy metrics for data analytics assessment and validation
Skills Data Analytics languages, tools and platforms	DSALANG01	R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)
	DSALANG02	Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
	DSABDA09	Kaggle competition, resources and community platform

Dimension 1 Competence Group	DSDA	Data Science Analytics
Dimension 2 Competence	DSDA01	Effectively use variety of data analytics techniques, such as Machine Learning (including supervised, unsupervised, semi-supervised learning), Data Mining, Prescriptive and Predictive Analytics, for complex data analysis through the whole data lifecycle
Dimension 3 Proficiency	Level 1 (Entry/Associate)	Level 1 (Professional)	Level 1 (Expert)
	Understand and be able to select an approach to analyzing asets. e understanding form statistical testing, explain gnificance.	Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction	Develop and plan required data analytics for organizational tasks, including: evaluating requirements and specifications of problems to recommend possible analytics-based solutions
D	Knowledge unit definition	
	Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others	
	Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)	
	Machine Learning (reinforced): Q-Learning, TD-Learning, Genetic Algorithms)	
	Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)	
	Predictive Analytics	
	Prescriptive Analytics	
	Data preparation and pre-processing	
	Performance and accuracy metrics	
	Skills definition	
	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)	
	Use Data Mining techniques	
	Apply Predictive Analytics methods	
	Apply Prescriptive Analytics methods	
	Use Graph Data Analytics for organisational network analysis, customer relations, other tasks	
	R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)	
	Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)	
	SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)	
	NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)	
	Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)	
	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)	
	Real time and streaming analytics systems (Flume, Kafka, Storm)	
	Kaggle competition, resources and community platform	
	Git versioning system as a general platform for software development	

Group 5: Soft skills and Workplace skills

- Data Science professional skills: Thinking and Acting like Data Scientist
- 21st Century Skills: Personal, inter-personal communication, team work, professional network
- Data Scientist and Subject Domain Specialist

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 of 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)

21st Century Skills (DARE & BHEF & EDISON - Updated)

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

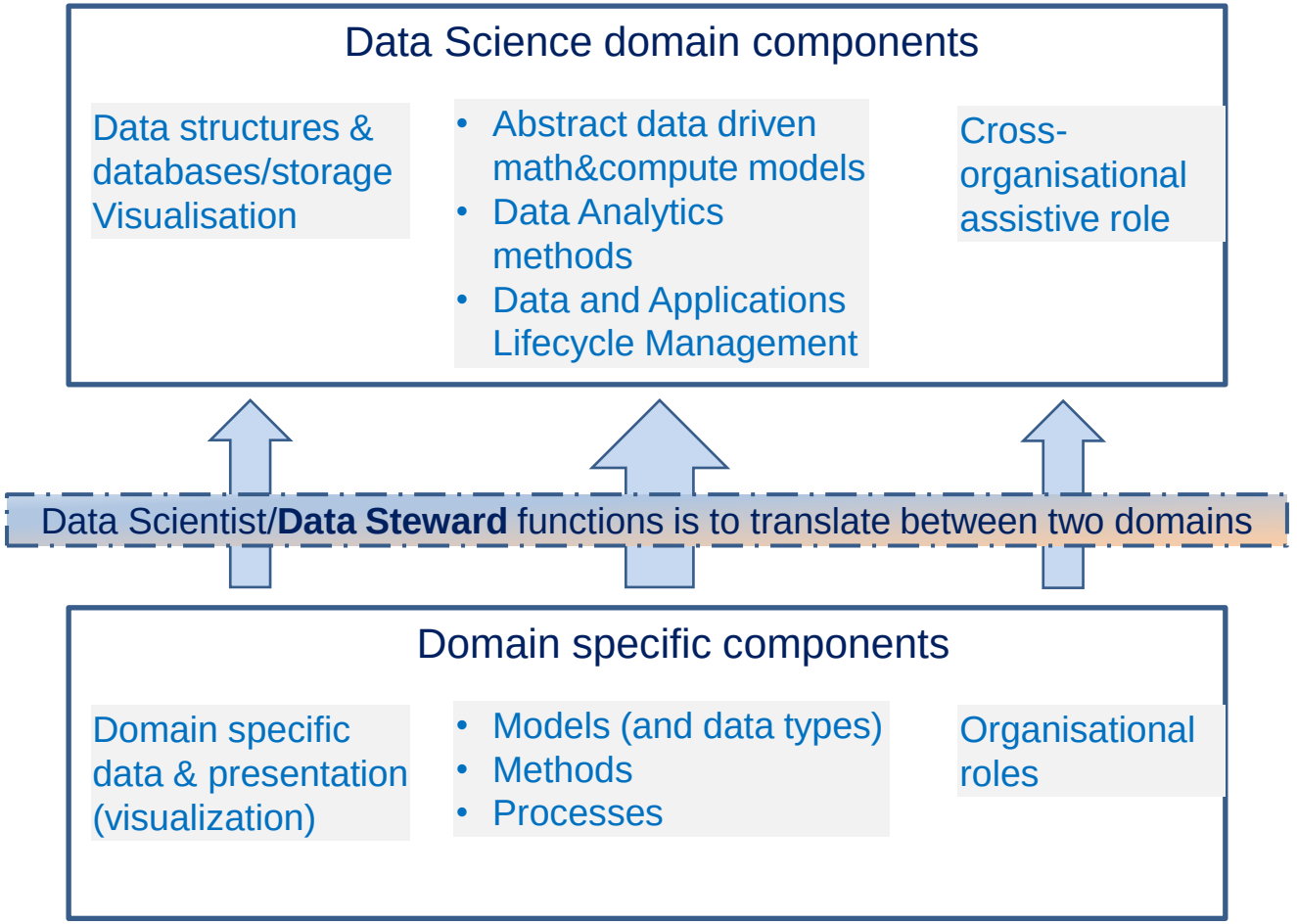
6. **Business Fundamentals:** Having fundamental knowledge of the organization and the industry
7. **Customer Focus:** Actively look for ways to identify market demands and meet customer or client needs
8. **Working with Tools & Technology:** Selecting, using, and maintaining tools and technology to 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
12. Green (environmental Sustainability) awareness in computation and data storage

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



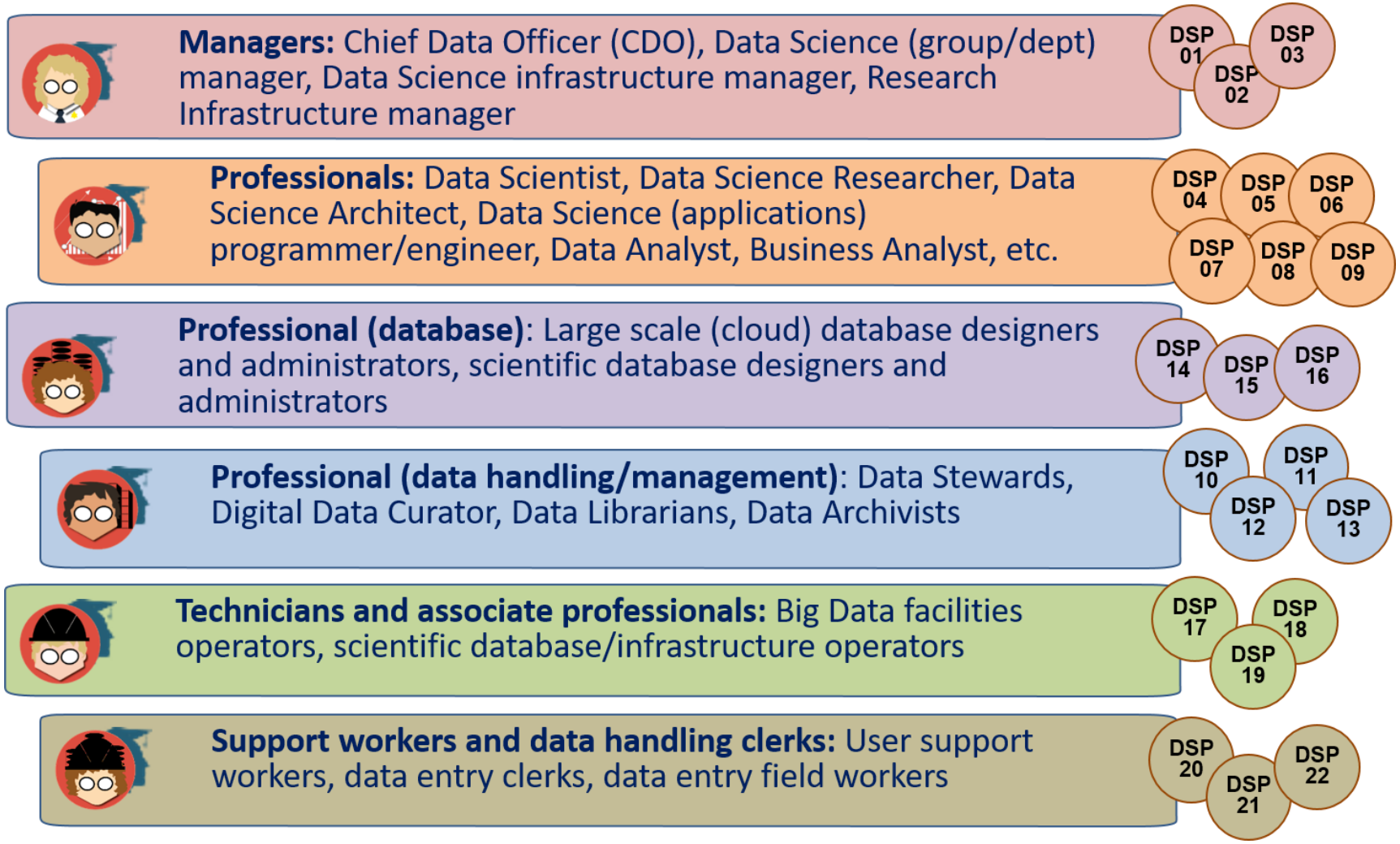
Data Scientist role is to maintain the Data Value Chain (domain specific):

- Data Integration => Organisation/Process/Business Optimisation => **Innovation**

Practical Application of the CF-DS

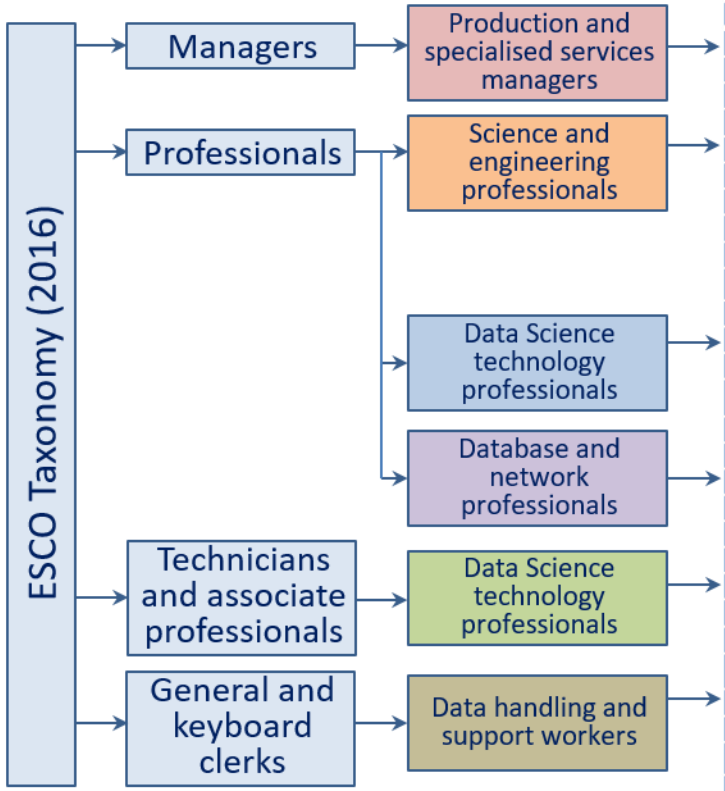
- Basis for the definition of the Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS)
  - CF-DS => Learning Outcomes (MC-DS) => Knowledge Areas (DS-BoK)
  - CF-DS => Data Science taxonomy of scientific subjects and vocabulary
- Data Science professional profiles definition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



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

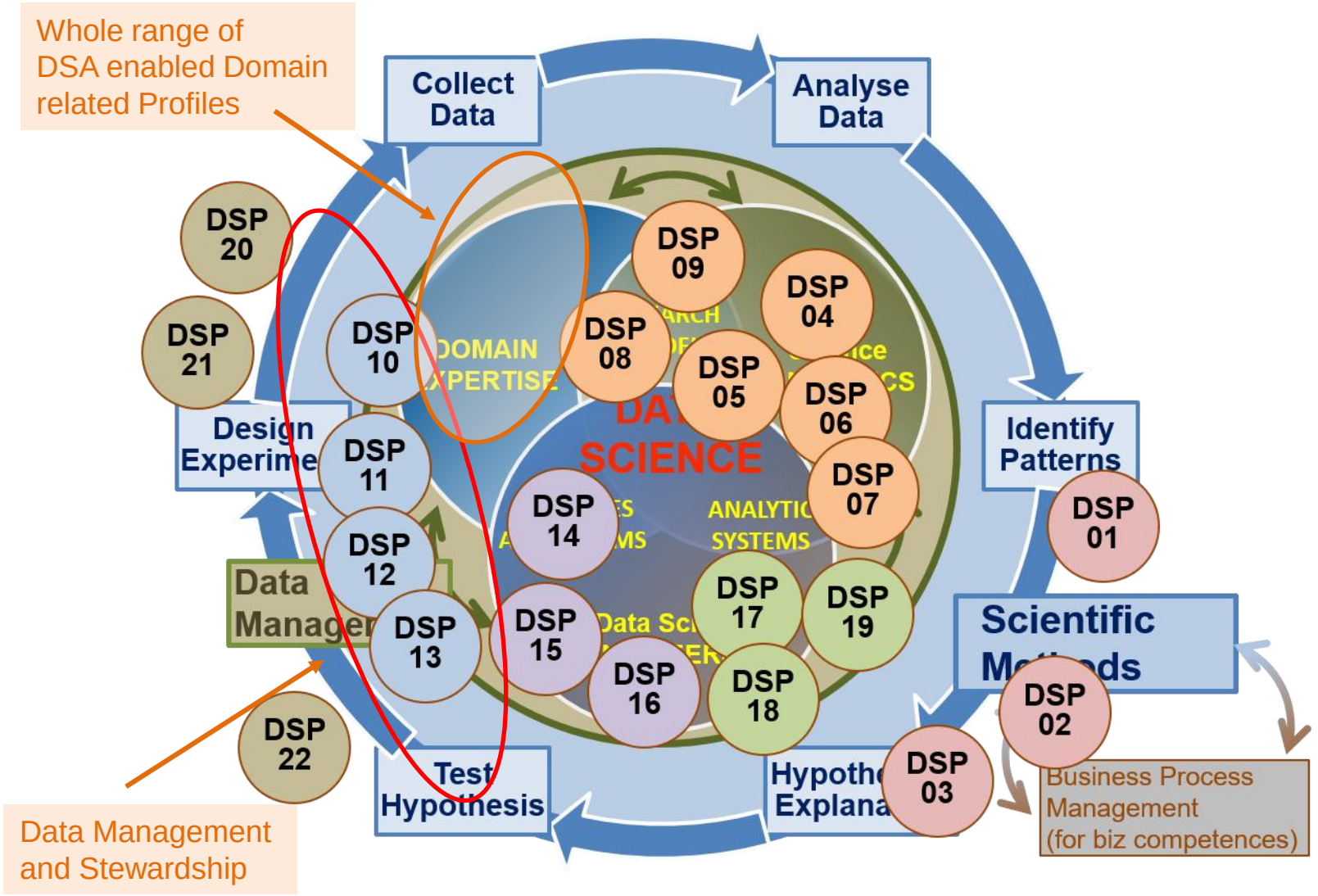
DSP Profiles mapping to ESCO Taxonomy High Level Groups



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

CF-DS and Data Science Professional Profiles



Example DS Professional Profile Definition (compliant with CWA)

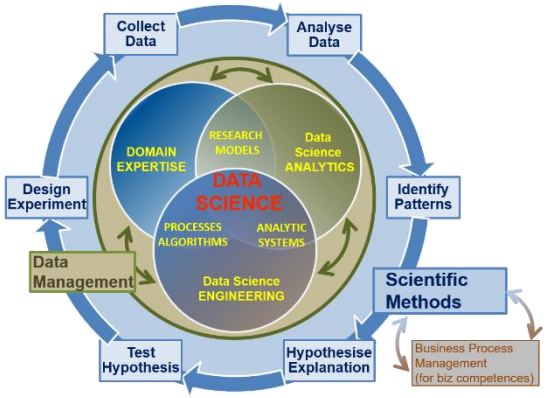
Profile title	**Gives a commonly used name to a profile. TEMPLATE**		
Summary statement	**Indicates the main purpose of the profile.** **The purpose is to present to stakeholders and users a brief, concise understanding of the specified ICT Profile. It should be understandable by ICT professionals, ICT managers and Human Resource personnel. It should provide a statement of the job's main activity.**		
Mission	**Describes the rationale of the profile.** **The purpose is to specify the designated job role defined in the ICT Profile.**		
Deliverables	Accountable (A)	Responsible (R)	**Contributor (C)**
	Specifies the Profile by key deliverables. **The purpose is to illuminate the ICT Profiles and to explain relevance including the perspective from a non-ICT point of view.**		
Main task/s	**Provides a list of typical tasks to be performed by the profile.** **A task is an action taken to achieve a result within a broadly defined context. Tasks may be associated with deadlines, resources, goals, specifications and/or the expected results.**		
e-CF competences assigned	**Provides a list of necessary competences (from the e-CF) to carry out the mission.** **Must include 1 up to 5 competences.** **Level assignment is important. Can be (usually) 1 or (maximum) 2 levels.**		
KPI Area	**Based upon KPIs (Key Performance Indicators) KPI area is a more generic indicator, congruent with the overall profile granularity level. It is deployed to add depth to the mission.** **Not prescriptive. Non-specific measurements. Use general examples.** **The principle is to provide KPI areas (which are stable, general and long lasting) providing users with an inspiration to enable development of specific KPI's for specific roles** **Must be related to the key deliverables in order to measure them.**		

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

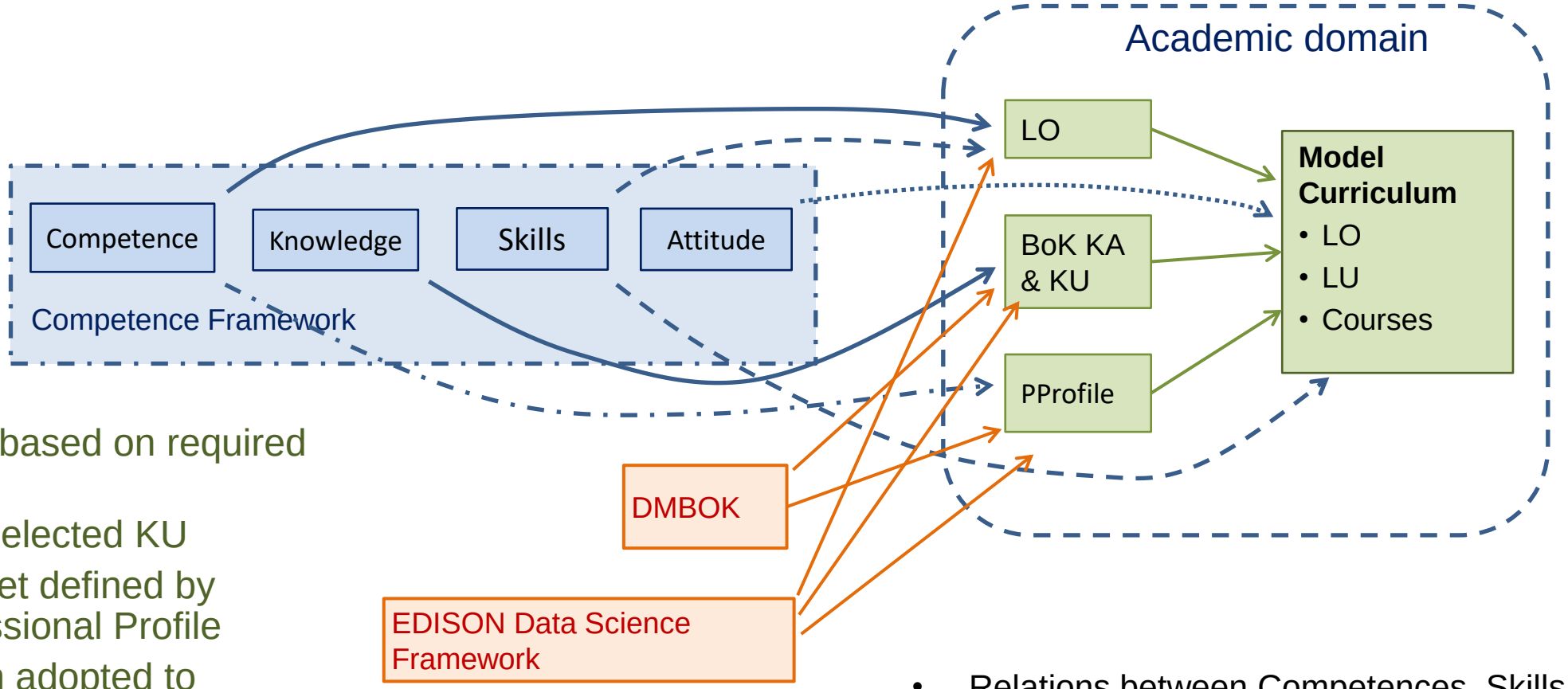
Data Science Body of Knowledge (1)

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

How to use CF-DSP and DSP-BoK for curriculum Design



- LO are defined based on required competences
- LU defined by selected KU
- Curriculum target defined by intended Professional Profile
- Final curriculum adopted to available resources
- Look *Adoption Handbook* for suggested courses

- Relations between Competences, Skills, Knowledge/BoK, Professional Profiles
- Mapping between Competence elements and Academic domain elements

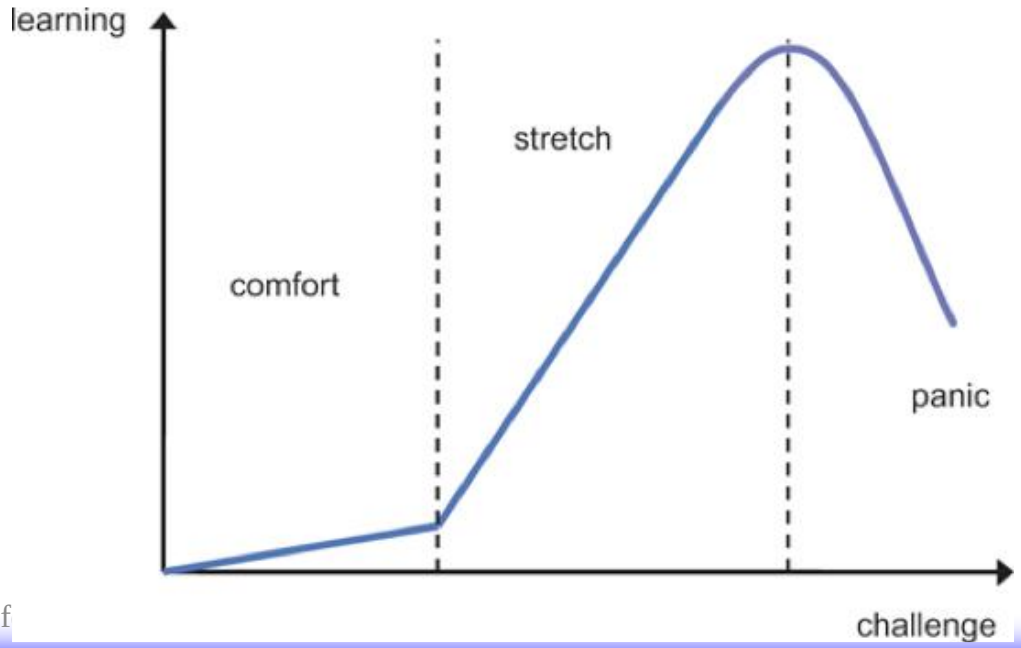
Data Science Model Curriculum (MC-DS)

Data Science Model Curriculum includes

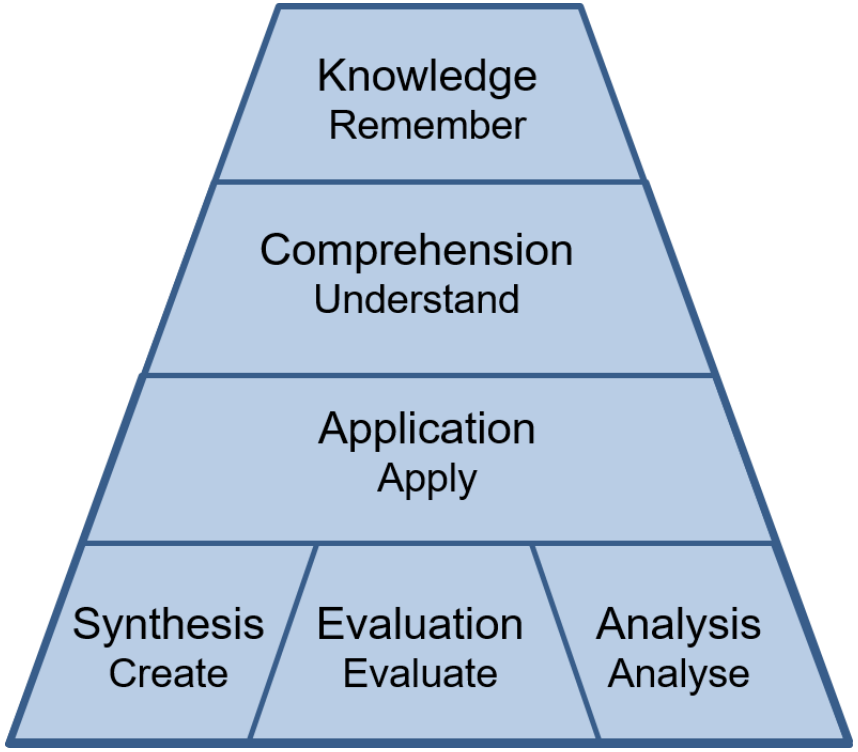
- Learning Outcomes (LO) definition based on CF-DS
  - LOs are defined for CF-DS competence groups and for all enumerated competences
  - Knowledge levels: Familiarity, Usage, Assessment (based in Bloom's Taxonomy)
- LOs mapping to Learning Units (LU)
  - LUs are based on CCS(2012) and universities best 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)

Learning methods and learning models

- Bloom's Taxonomy and Cognitive learning activities
  - BT application areas and limitations
- Constructive Alignment and Intended Learning Outcome (ILO)
  - ILO is formulated from the student perspective
  - Outcome Based Learning (OBL)
- Other education technologies for teaching in fast technology changing world
  - Project Based Learning (PBL)
  - Flipped classroom
  - Activating teaching and activating strategies

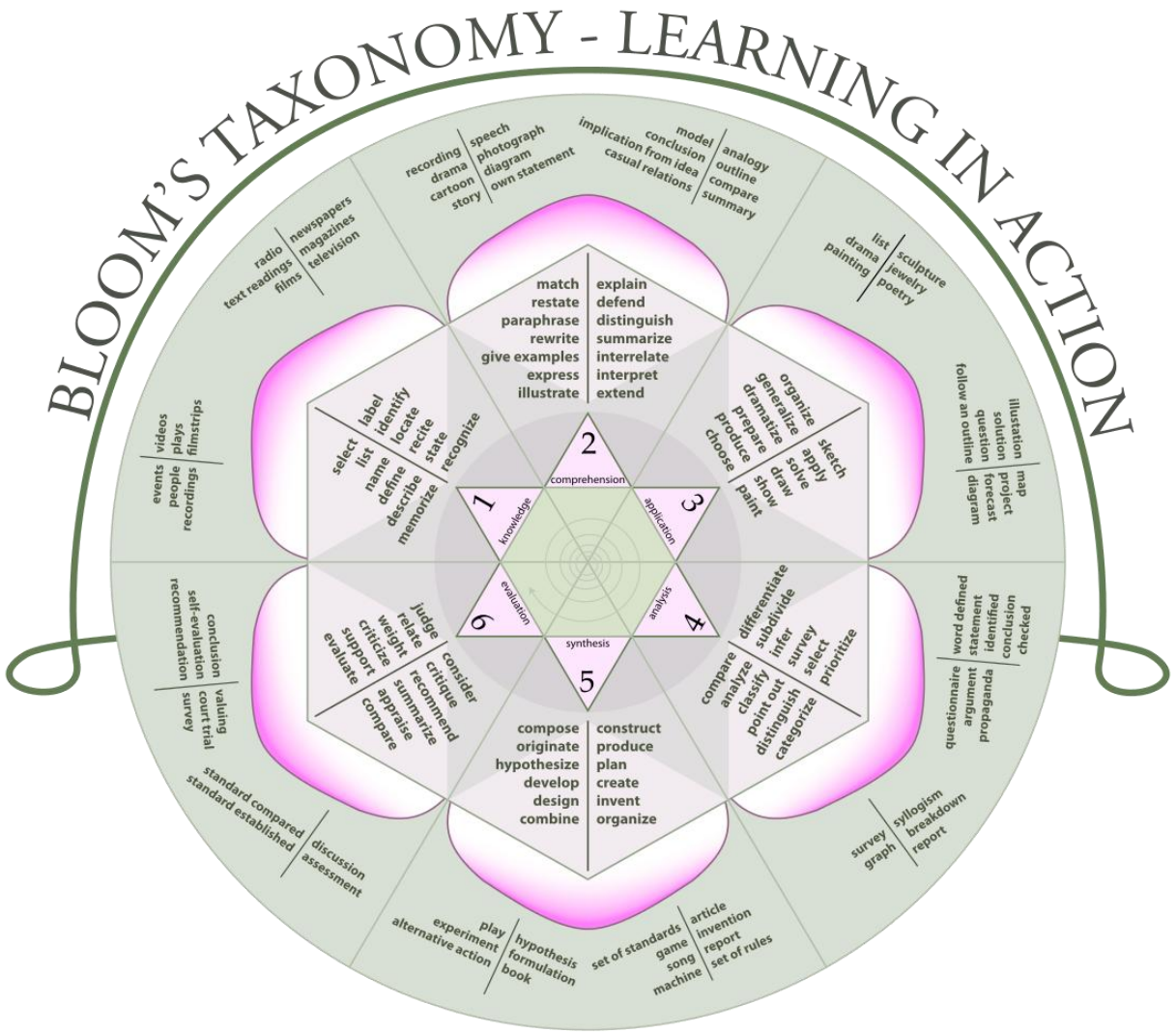


Bloom's Taxonomy and Knowledge Levels for MC-DS



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

Extended Bloom's Taxonomy



Consolidated presentation of learning levels, action verbs, and **associated learning instruments**

[ref] Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing, Abridged Edition. Boston, MA: Allyn and Bacon.

Bloom's Taxonomy – Cognitive Activities

Knowledge

Exhibit memory of previously learned materials by recalling facts, terms, basic concepts and answers

- Knowledge of specifics - terminology, specific facts
- Knowledge of ways and means of dealing with specifics - conventions, trends and sequences, classifications and categories, criteria, methodology
- Knowledge of the universals and abstractions in a field - principles and generalizations, theories and structures
- **Questions like: What are the main benefits of implementing Big Data and data analytics methods for organisation?**

Comprehension

Demonstrate understanding of facts and ideas by organizing, comparing, translating, interpreting, describing, and stating the main ideas

- Translation, Interpretation, Extrapolation
- **Questions like: Compare the business and operational models of private clouds and hybrid clouds.**

Application

Using new knowledge. Solve problems in new situations by applying acquired knowledge, facts, techniques and rules in a different way

- **Questions like: What data analytics methods should be applied for specific data types analysis or for specific business processes and activities? Which Big Data services architecture is best suited for medium size research organisation or company, and why??**

Analysis

Examine and break information into parts by identifying motives or causes. Make inferences and find evidence to support generalizations

- Analysis of elements, relationships, organizational principles
- Questions like: What data analytics methods and services are required to support typical business processes of a web trading company? Give suggestions how these services can be implemented with the selected data analytics platform, including on-premises or outsourced to cloud. Provide references to support your statements.

Synthesis

Compile information together in a different way by combining elements in a new pattern or proposing alternative solutions

- Production of a unique communication, a plan, or proposed set of operations, derivation of a set of abstract relations
- Questions like: Describe the main steps and tasks for implementing data analytics and data management services for an example company or research organisation? What services and data analytics can be moved to clouds and which will remain at the enterprise premises and run by company's personnel?

Evaluation

Present and defend opinions by making judgments about information, validity of ideas or quality of work based on a set of criteria

- Judgments in terms of internal evidence or external criteria
- **Questions like: Do you think that implementing Agile Data Driven Enterprise model creates benefits for enterprises, short term and long term?**

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 (defined based on Bloom's Taxonomy)

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

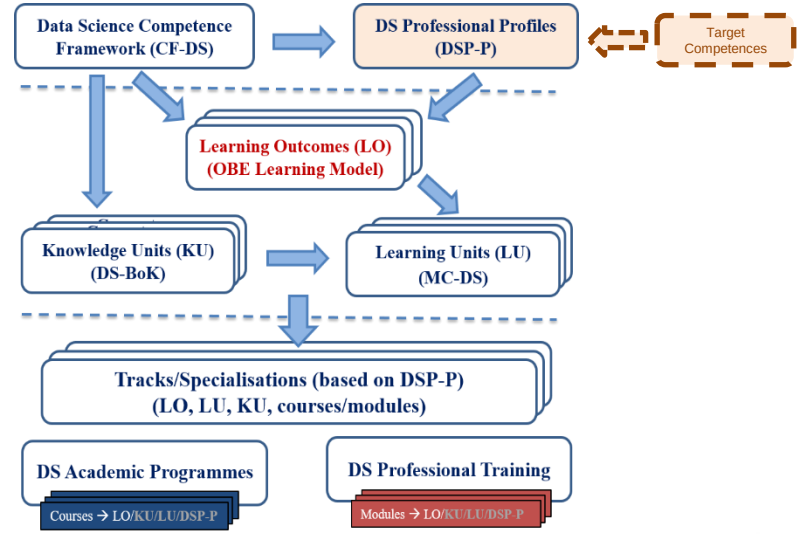
(14) Open Science, Open Data, Open Access, ORCID

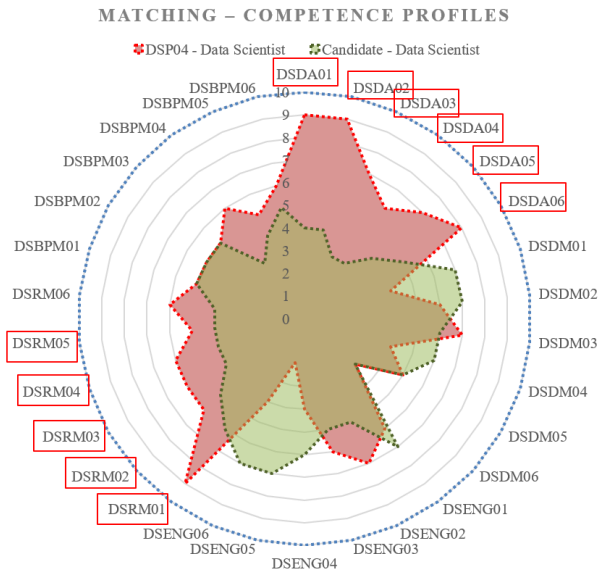
(15) Responsible data use

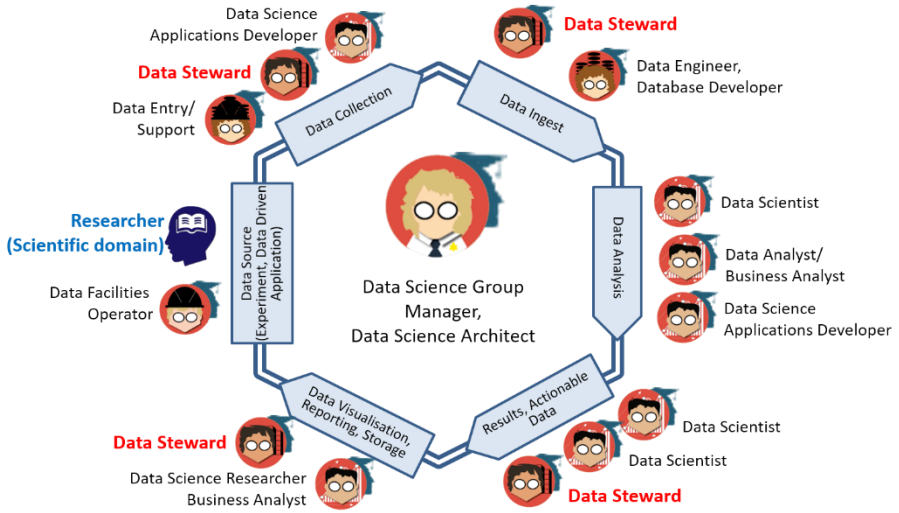
- Highlighted in red: Considered (Research) Data Management literacy (minimum required knowledge)

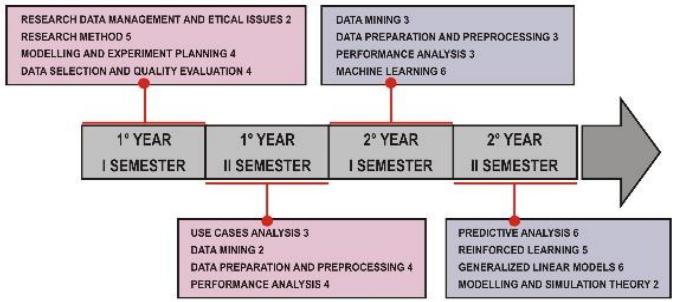
How to use CF-DSP and DSP-BoK

- Customised Curriculum Design
- Competences – Vacancies assessment
- Data Science/Data Steward Team building

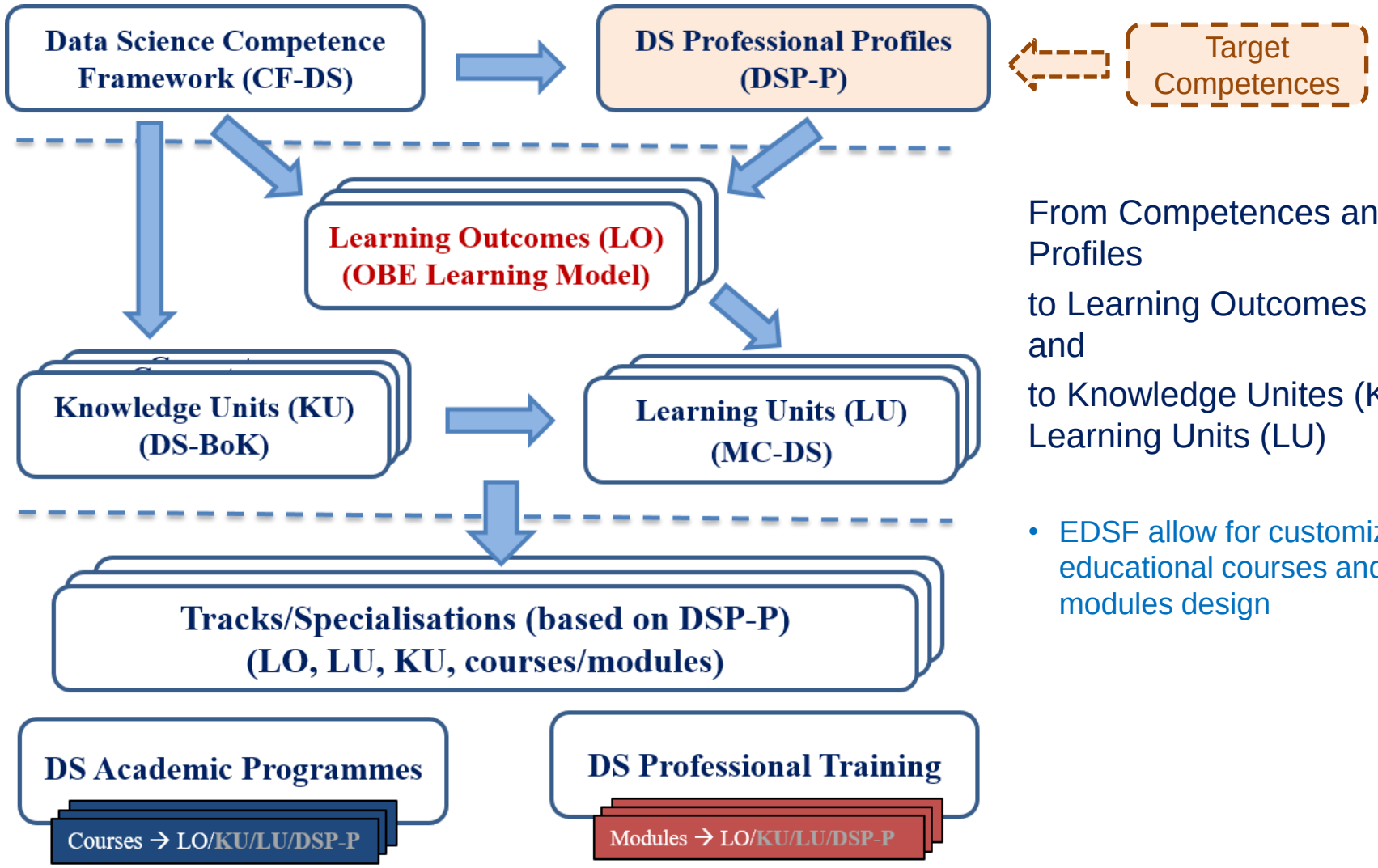








Outcome Based Educations and Training Model: Addressing target competences for the profession



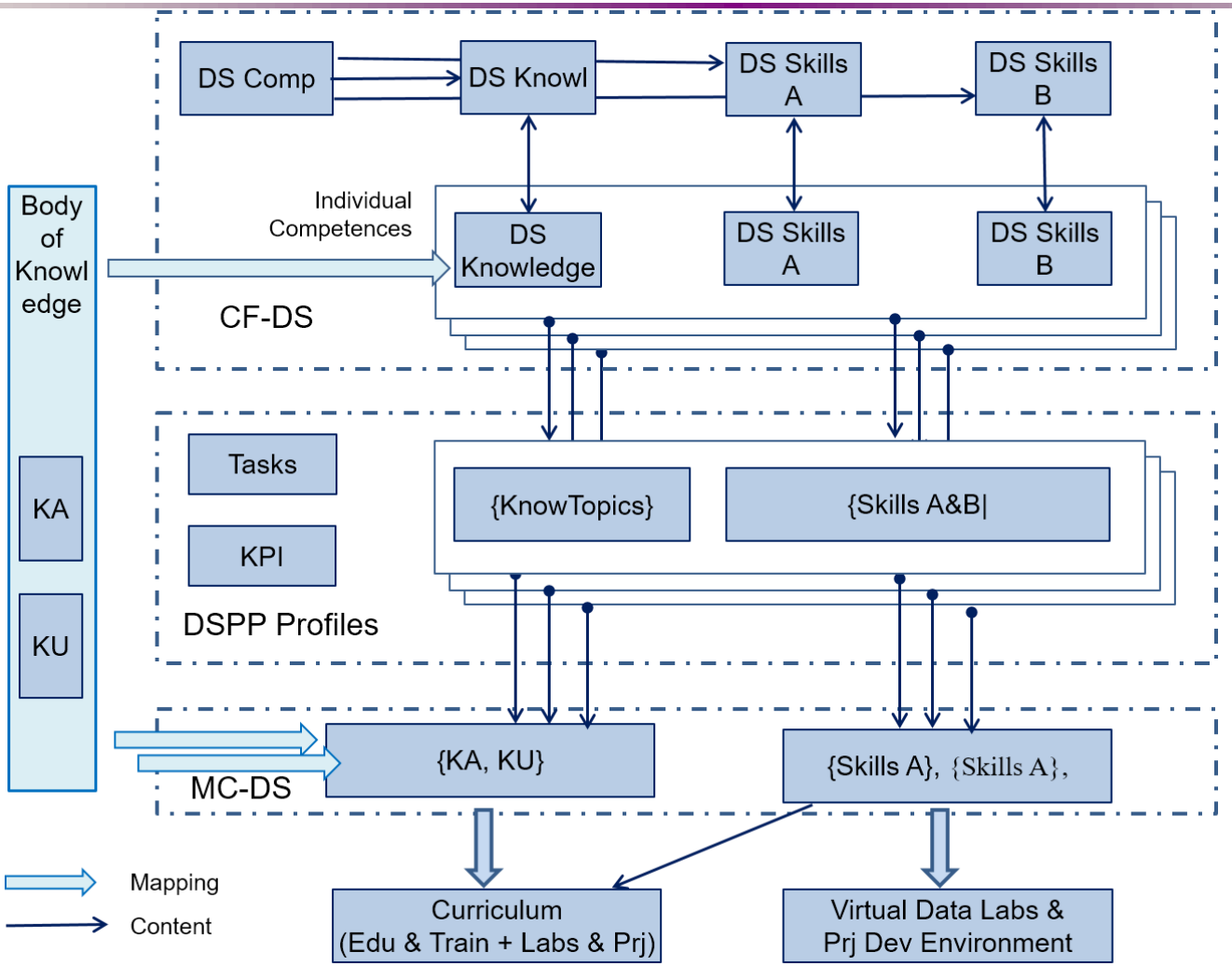
From Competences and DSP Profiles

to Learning Outcomes (LO) and

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

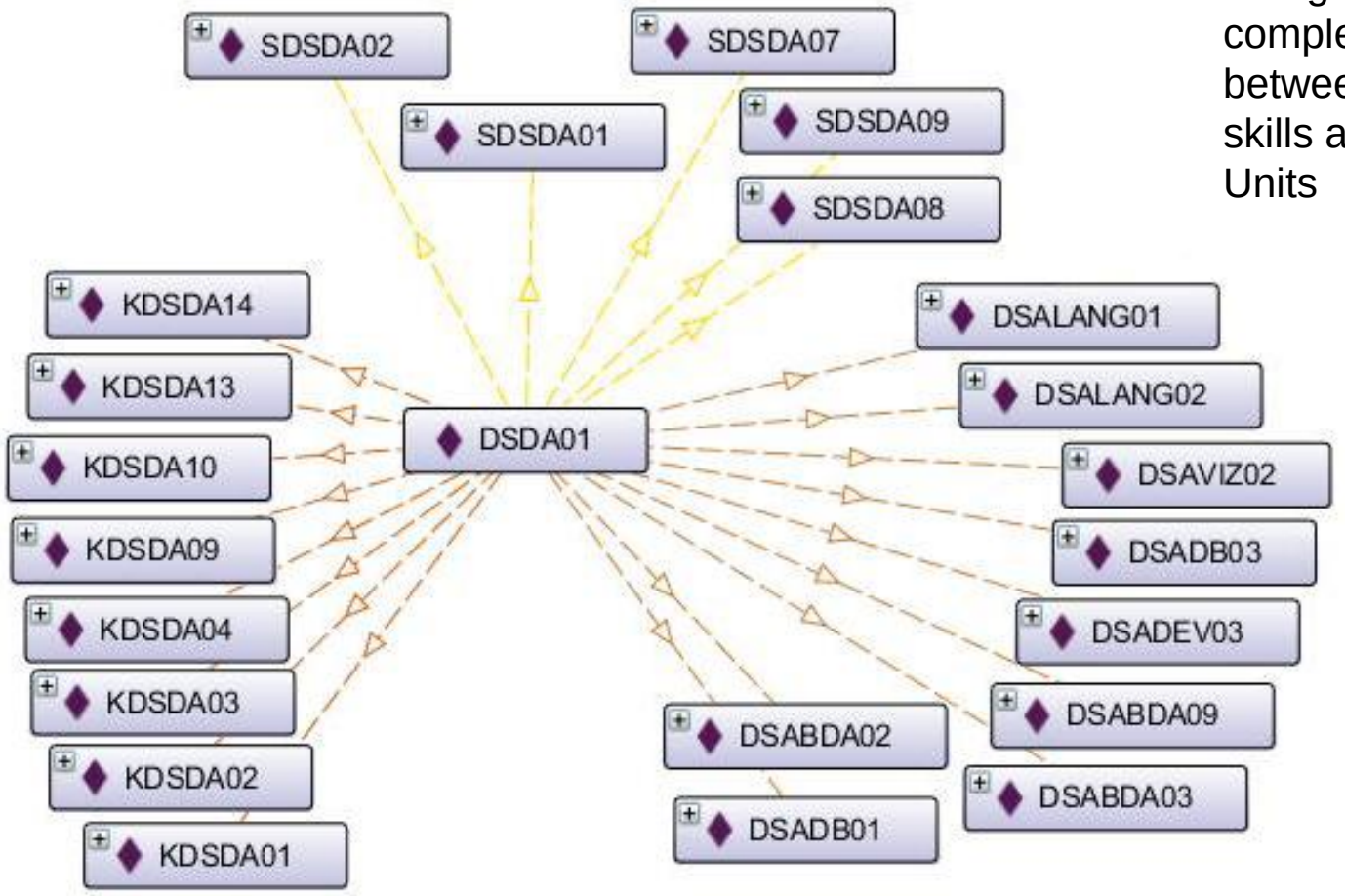
- EDSF allow for customized educational courses and training modules design

EDSF Data Model and API



- EDSF API provides access to all EDSF functionality

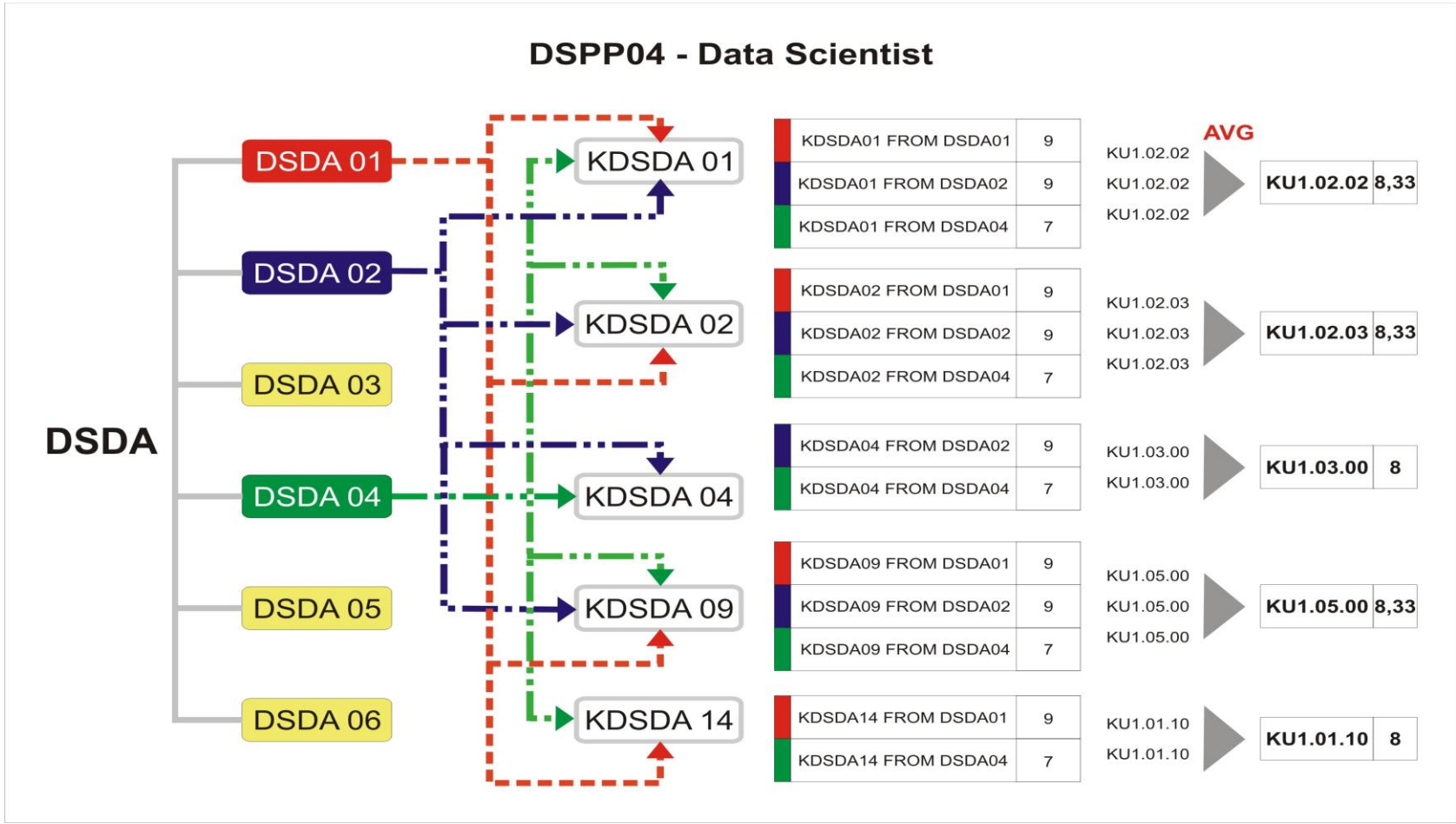
Example DSDA01 Competence and its properties



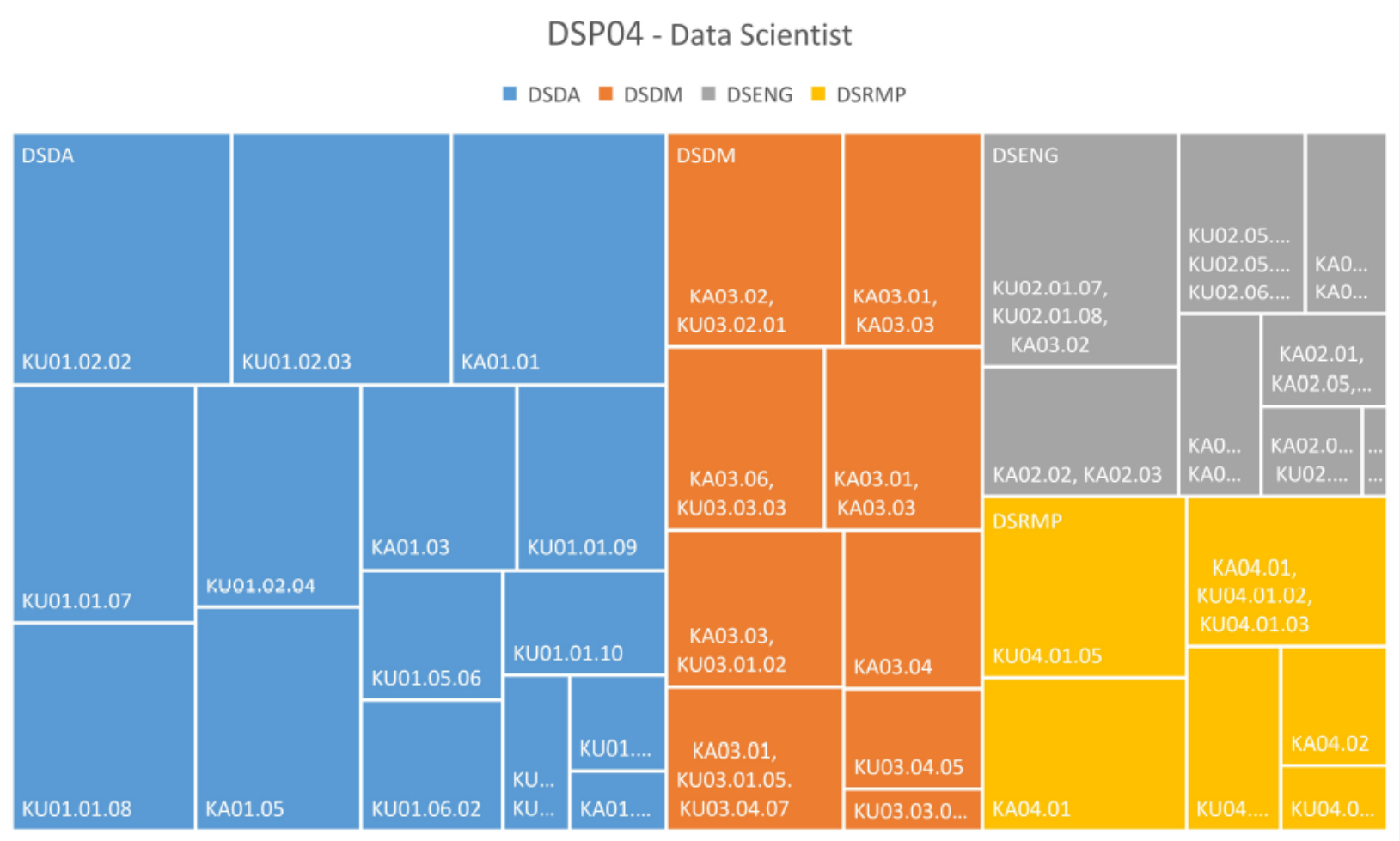
- Using ontology to manage all complexity of relations between Knowledge topics, skills and BoK Knowledge Units

- KDSDA – Knowledge topics
- SDSDA – Skills related to DSDA01
- DSAlang, DSAdb, DSAbda – skills practical knowledge

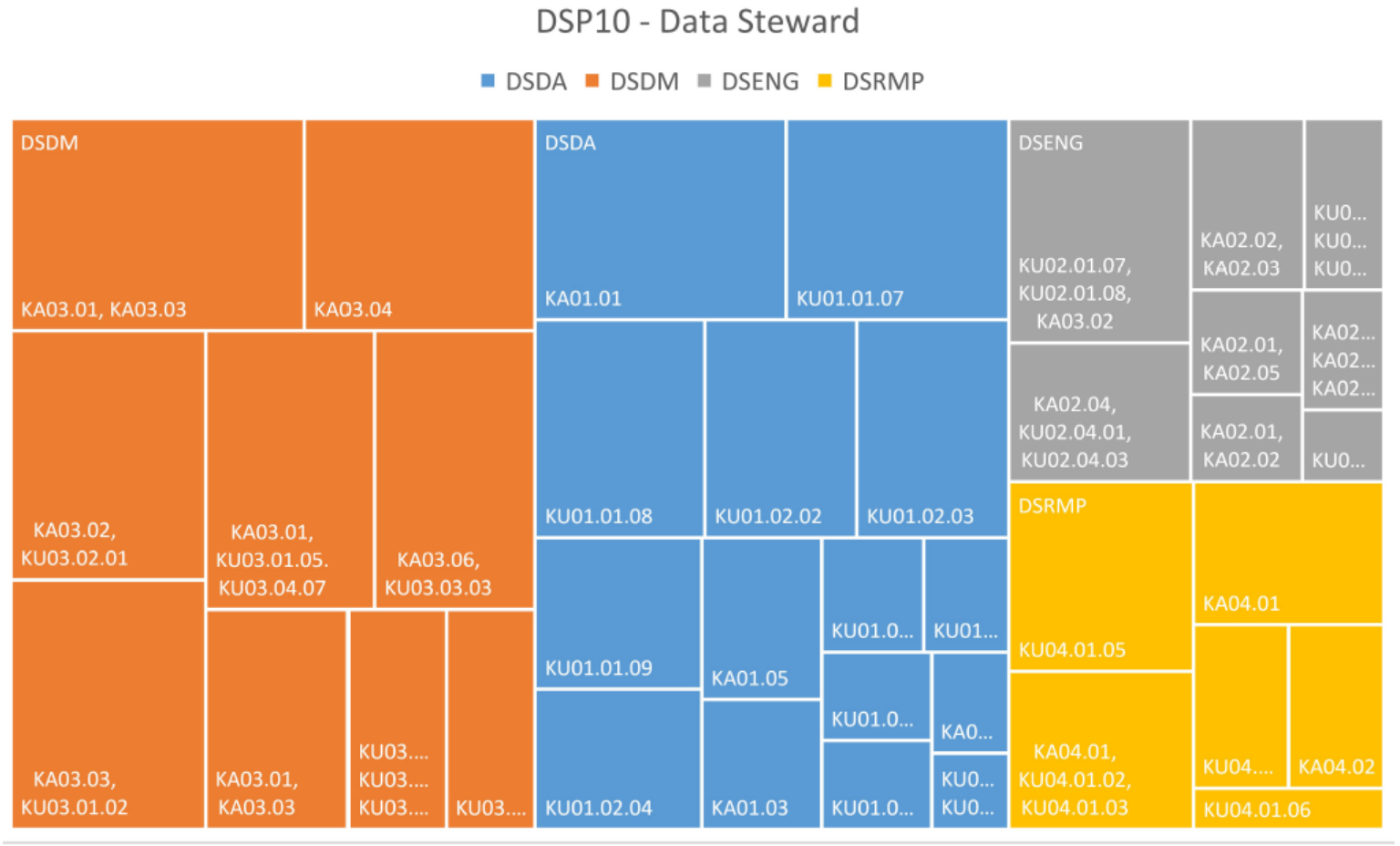
Extracting required Knowledge Units from EDSF ontology for DSPP04 – Data Scientist



DSP04 – Data Scientist MC structure



DSP10 – Data Steward MC structure



DSP04 Data Scientist – Required practical skills and Hands-on labs

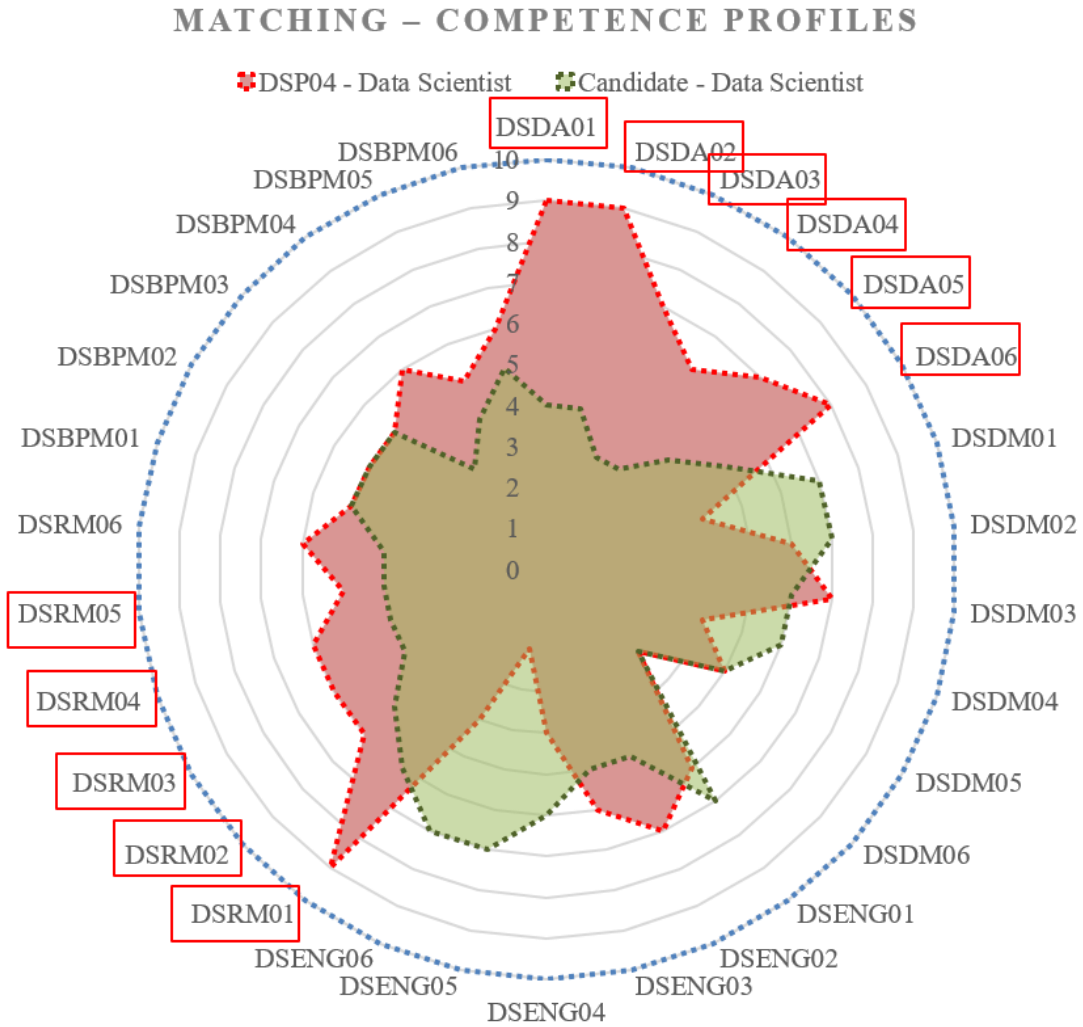
Data Science curriculum should include the following elements to achieve necessary skills Type B:

- Python (or R) and corresponding data analytics libraries
- NoSQL and SQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, MS SQL, My SQL, PostgreSQL, etc.)
- Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
- Real time and streaming analytics systems (Flume, Kafka, Storm)
- Kaggle competition, resources and community platform, including rich data sets, forum and computing resources
- Visualisation software (D3.js, Processing, Tableau, Julia, Raphael, etc.)
- Web API management and web scrapping
- Git versioning system as a general platform for software development
- Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others
- Cloud based Big Data and data analytics platforms and services, including large scale storage systems
  - Essential for workplace alignment

Competences assessment and Team building

- Data Science competences assessment (benchmarking)
- Data Science team building

Individual Competences Benchmarking

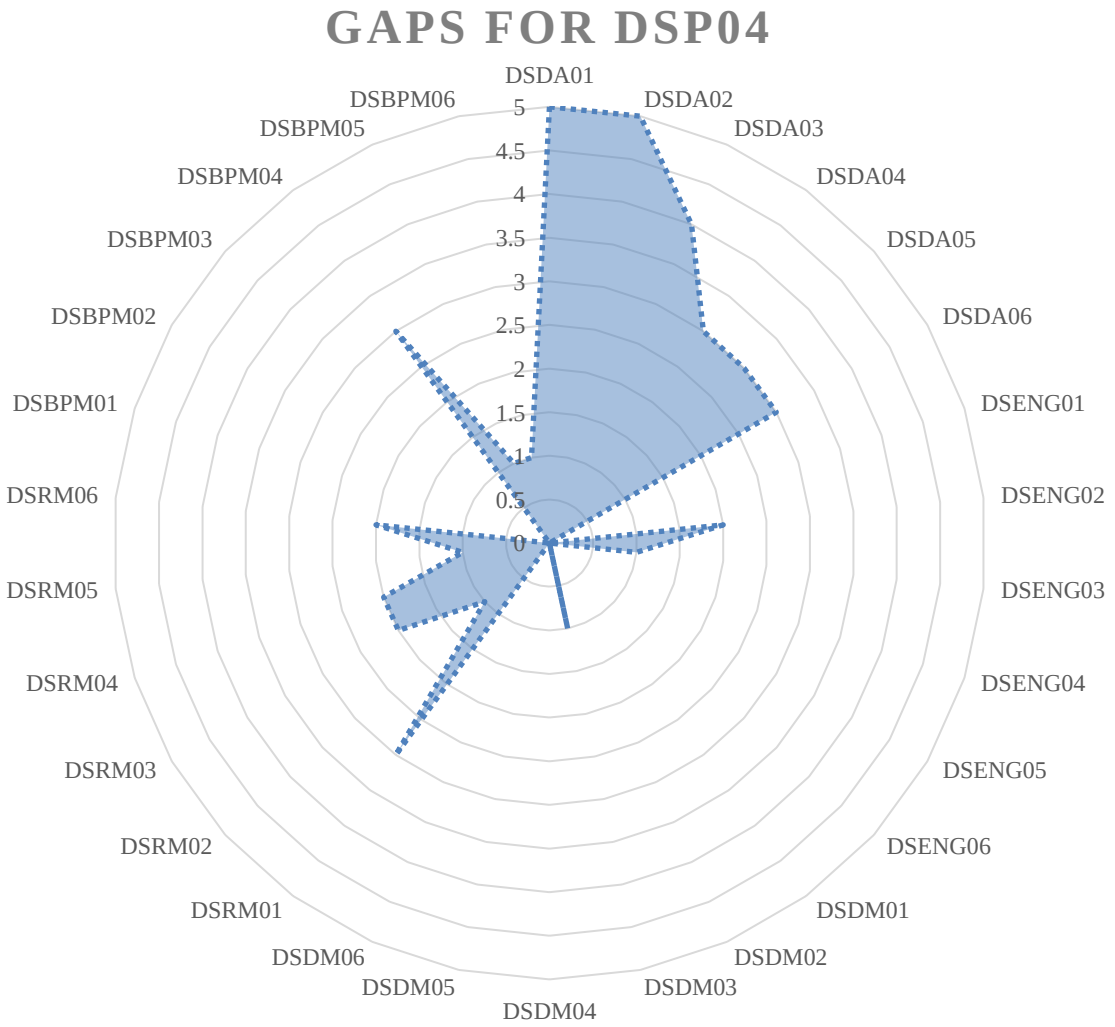


Individual Education/Training Path based on Competence benchmarking

- 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.

Competence/Knowledge gap -> Suggested LUs/courses



Recommended courses DSDA

- Statistical Methods
- Machine Learning
- Predictive and Quantitative analytics
- Graph Data Analysis
- Data preparation and preprocessing
- Performance Analysis

Recommended courses DSRM

- Research Methods and Project Management

Mapping to career path: Mastery Levels against Competences Relevance

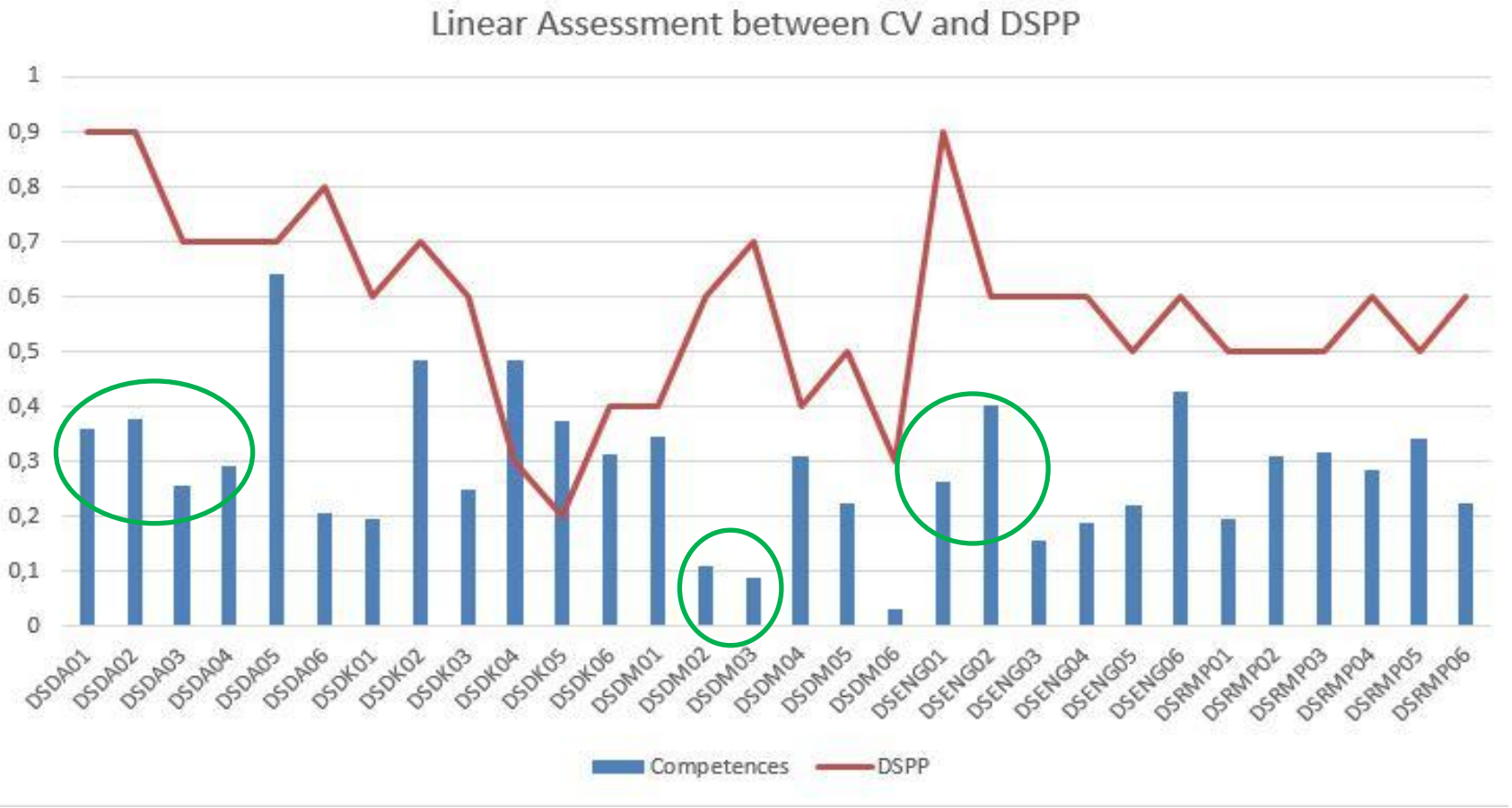
Mastery levels defined using workplace terminology that can be easy mapped to mastery levels defined in MC-DS:

- A - Awareness
  - 1) Understand Terminology
  - 2) Understand Principles
  - 3) Apply principles
  - 4) Understand Methods
- U - Use/Application
  - 5) Apply basics
  - 6) Supervised use
  - 7) Unsupervised Use
- P - Professional/Expert
  - 8) Development of applications using wide range of technologies
  - 9) Supervise project development, team of professionals

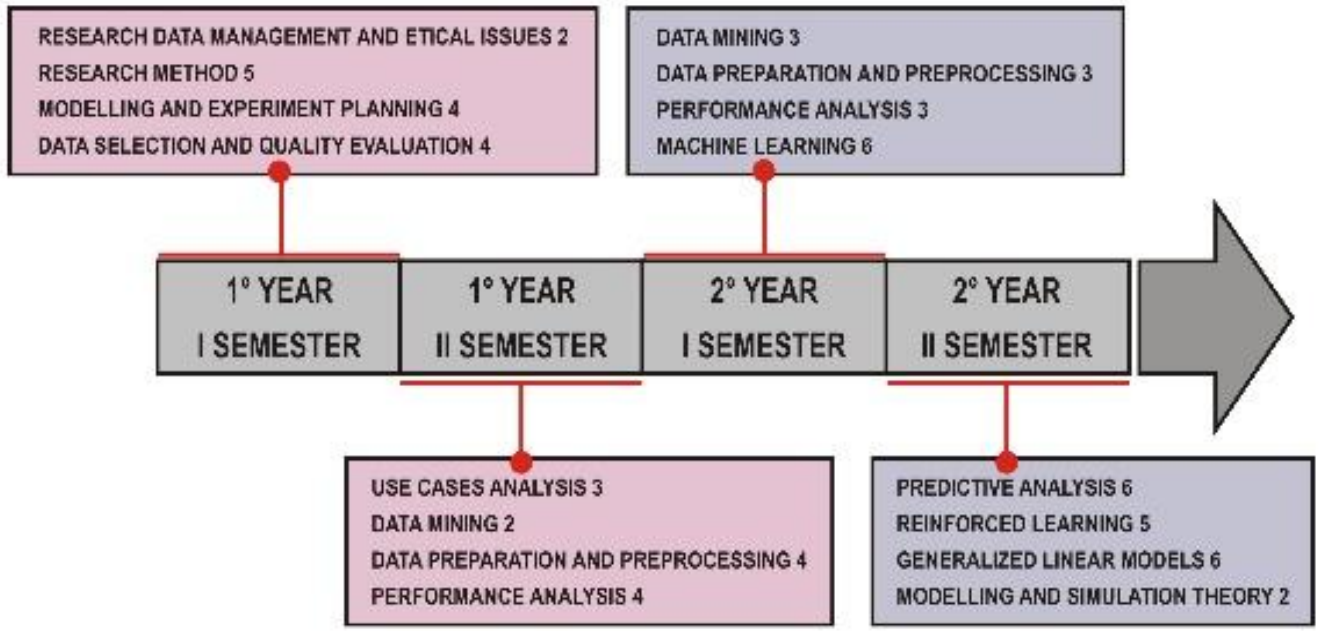
User/Professional level: Measurable competences and knowledge

Professional/Expert level: Peer-reviewed competences and skills

From Competence gap to Proficiency Level and Curriculum Timing – In progress



Example curriculum planning: Based on implied courses duration and DSPP profile proficiency levels



- Planning on sequence and duration of courses for maximum learning outcome
- Importance of splitting core courses over 2 semesters
- Theory and practice oriented courses
- Importance of pre-requisite knowledge:
  - Statistics for Data Scientists
  - Organisational management for Data Stewards

Managing Data Science Teams - Roles

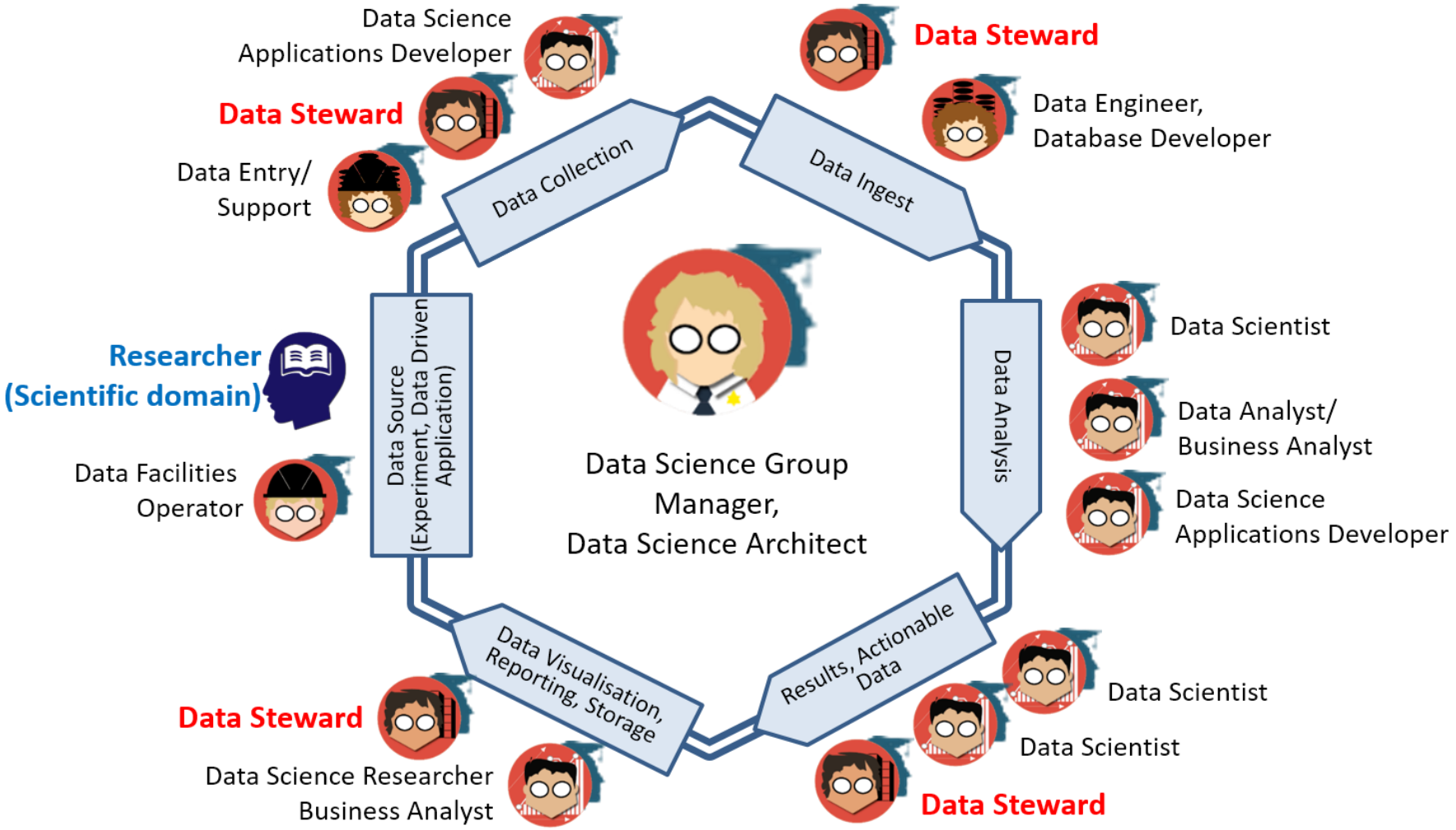
- Data science team consists of Data Scientists, Data Architects/Engineers, Machine Learning engineers and Software Developers
- **Data Scientist** solves business problems using machine learning and data mining techniques. Data Scientists are also responsible for using statistical methods, processes, and algorithms to extract insights from data. The tasks include data pre-processing, analyzing , perform experiments on it, visualizing it and communicate those result.
- **Machine Learning Engineer** combines software engineering and modeling skills. Everything that goes into training, monitoring, and maintaining a model is ML engineer's job.
- **Data Architect/Engineer-** Responsible for implement, test, and maintain infrastructural components for big data and large distributed systems.
- **Software Developer** is responsible for taking a deployed model and make it ready to be served through REST API's and may involve some front-end interface as well, so a software developer helps in all these tasks.
- **Data Annotation and Quality Assurance(QA)-** Data is Key to success of any data science team. Having a well trained Data labeling team can provide significant value particularly during the iterative machine learning model testing and validation stages. Machine learning is an iterative process. You need to validate the model predictions and also need to prepare new datasets and enrich existing datasets to improve your algorithm's results. You can either hire or outsource data annotation but challenge remains to have a consistency in labeling data and validating results.
- **Research Scientist-** If your team works on some core AI domain like Conversational AI, Computer Vision, Robotics, Reinforcement Learning, Graphical Models etc. you might need to hire someone with a PhD or core research background.
- **Data Science Manager-** Responsible for recruiting and building data science teams, showcasing the team capabilities, interfacing with senior management, develop the process that the team can follow, help in team communications and to keep things moving.

[ref https://towardsdatascience.com/building-and-managing-data-science-teams-77ba4f43bc58

Managing Data Science Teams - Process

1. **Growing the team:** Initially you may need a small team which mostly works on some analysis or come up with some ideas which you can pitch up to the senior management. But you will soon realize that to build the idea into a product your team needs to have many other skills. The aim should be to grow the data science team into a full product team responsible for designing, implementing, and maintaining products. As a product team, data science team could experiment, build, and add value directly to the company.
2. **Prioritize work:** I have seen that every now and then the team is flooded with requests for some analytics report or some other data crunching requests. These adhoc requests consume lot of time and it impacts the long term projects and other key deliverable. It's important prioritize work and assign right priority to these adhoc tasks. In our team,we created an adhoc requests backlog and added priority to these tasks. The team could then manage these urgent requests better without sacrificing the time towards important tasks.
3. **Data quality:** The first question is: Are you getting the right data? You may have plenty of data available, but the quality of that data isn't a given. To create, validate, and maintain production for high-performing machine learning models, you have to train and validate them using trusted, reliable data. You need to check both the **Accuracy** and **Quality** of the data. Accuracy in data labeling measures how close the labeling is to ground truth. Quality in data labeling is about accuracy across the overall dataset. Make sure that the work of all of your annotators look same and labeling is consistently accurate across your datasets.
4. **Tools :** Tools play an important role because they allow you to automate. You should use relevant tools to do heavy lifting jobs, running scripts to automate queries and processing data to save some time which can in turn be used to make the team more productive. Data science team is motivated by solving challenging problems. Automating repetitive weekly reports can help engineers to focus on some new challenging problems. In our team, we made a tool for labeling our data and exposed the tool to data annotation team. That really helped us to check for data consistency and share the work across different members with quick turn around time for the labeling task.
5. **Processes:** Data science team projects are research oriented or start with lot of research activities , it's difficult to predict how long it will take for them to finish. Also lot of activities like model building, data crunching are usually done by a single person, so traditional collaborative workflows don't fit. You have to identify an approach which works best for your team. Like in our case, we run a mix of Kanban and Scrum boards in JIRA. For Research Activities, Data Exploration/Analysis, Exploring ML models go for Kanban mode while as **Productization** of the models you can work as a Scrum team. So basically your Data scientists, Research Scientists and ML Engineers work mostly in Kanban mode where as Data Engineers, Software Engineers work in Scrum mode. Evaluate various options and see what best works for your team and projects.

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 Stewardship in Research and FAIR Principles

- FAIR Initiative by Dutch Techcentre for Life Science (DTLS) – Prof. Barend Mons
  - Supported by Germany, France, Spain, UK, USA
  - Part of Horizon 2020 Programme
- FAIR Principles for research data:
  Findable – Accessible – Interoperable - Reusable
- 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 the 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.

HLEG report on European Open Science Cloud
(October 2016)

Data Management and Governance (DMG) and Research Data Management (RDM)

- RDM curricula example

- DAMA Data Management Body of Knowledge (DMBOK)

Research Data Management Model Curriculum – Part of the EDISON Data Literacy Training

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

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

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

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

E. Hands on:

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

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

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

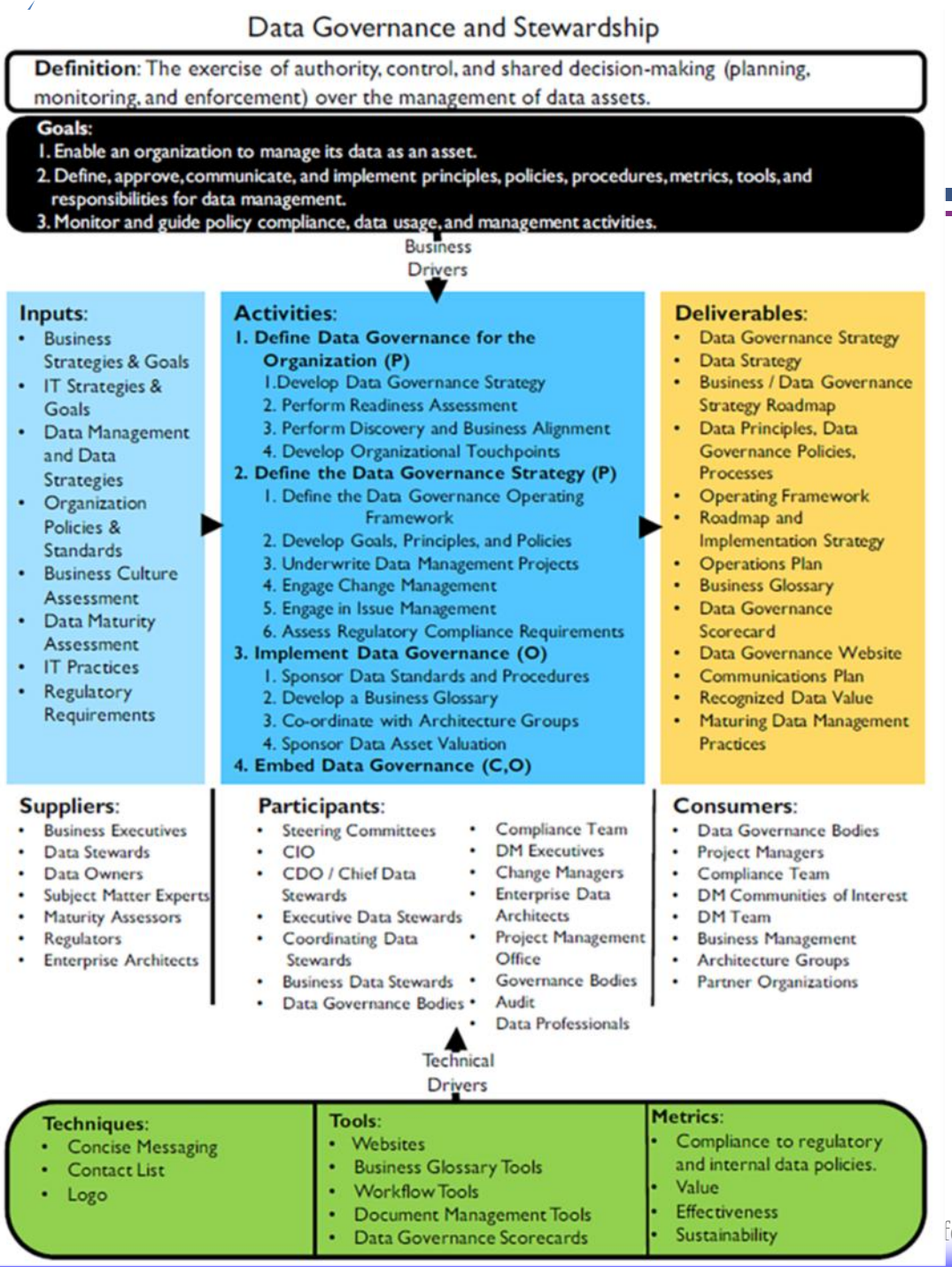
(13) Data Management Plan

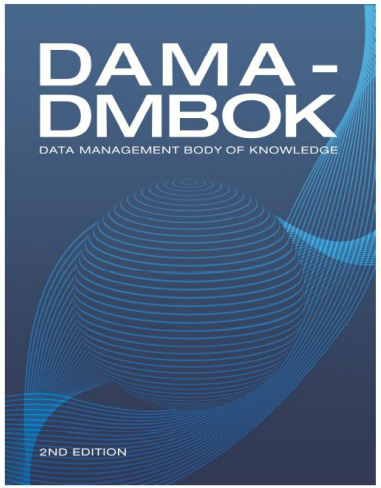
(14) Open Science, Open Data, Open Access, ORCID

(15) Responsible data use

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

DMBOK: Data Governance and Stewardship

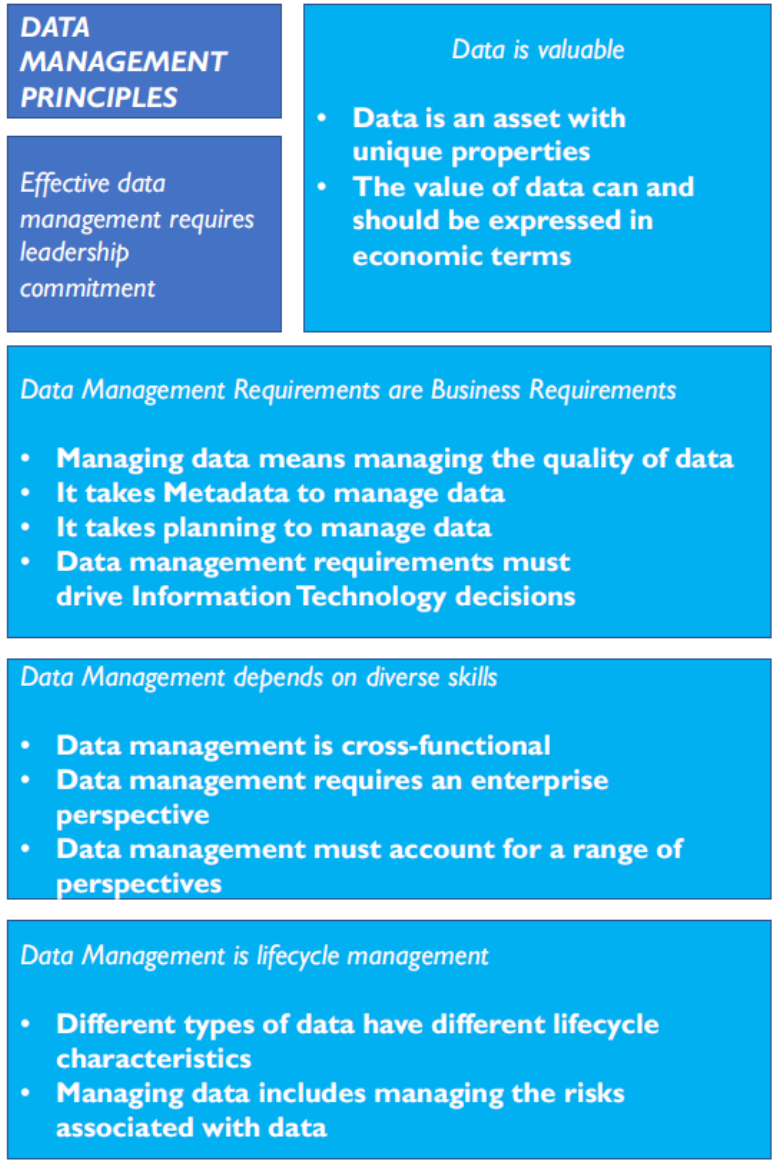




Scope of a Data Governance Programme

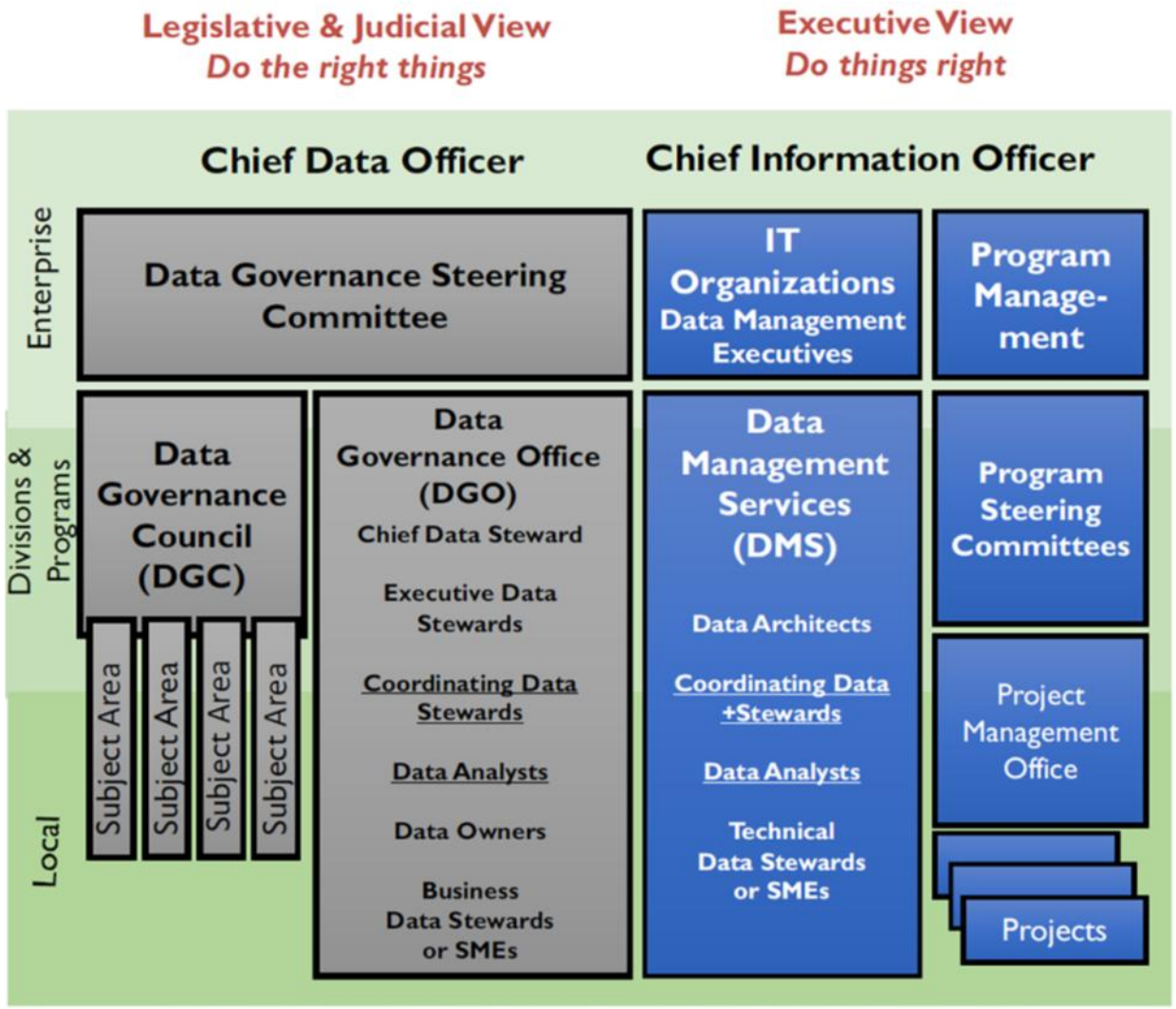
- Strategy
- Policy
- Standards and quality
- Oversight
- Compliance
- Issue management
- Data management projects
- Data asset valuation

DMBOK: Data Management Principles



- Data is an asset with unique properties
- The value of data can and should be expressed in economic terms
- Managing data means managing the quality of data
- It takes Metadata to manage data
- It takes planning to manage data
- Data management requirements must drive Information Technology decisions
- Data management is cross-functional; it requires a range of skills and expertise
- Data management requires an enterprise perspective
- Data management must account for a range of perspectives
- Data management is lifecycle management
- Different types of data have different lifecycle characteristics
- Managing data includes managing the risks associated with data
- Effective data management requires leadership commitment

DMBOK: Data Governance Organisation Parts



- Separation or governance responsibilities
- Multi-layer
- CDO
- CIO
- Councils

Data Stewardship

- **Creating and managing core Metadata**: Definition and management of business terminology, valid data values, and other critical Metadata.
- **Documenting rules and standards**: Definition/documentation of business rules, data standards, and data quality rules.
  - High quality data are often formulated in terms of rules rooted in the business processes that create or consume data.
  - Stewards help surface these rules and ensure their consistent use.
- **Managing data quality issues**: Stewards are often involved with the identification and resolution of data related issues or in facilitating the process of resolution.
- **Executing operational data governance activities**: Stewards are responsible for ensuring that, day-today and project-by-project, data governance policies and initiatives are adhered to. They should influence decisions to ensure that data is managed in ways that support the overall goals of the organization.

“Best Data Steward is not made but found” DMBOK1 (2009)

Ongoing and past activities and developments
https://github.com/EDISONcommunity/EDSF

- **EDISON Data Science Framework Release 3** – Published December 2018
- **EDISON Data Science Framework Release 4** – Published December 2022
  - Call for sponsorship
- **FAIRsFAIR project (2019-2022)**
  - Data Stewardship and FAIR data principles in the university curricula
  - Data Stewardship competences and Body of Knowledge definition
- **Industry digitalisation projects and data literacy skills training**
  - **MATES project funded by EU ERASMUS Programme -** EU Maritime industry digital transformation, data skills development and Data + Ocean literacy
  - Skills for SME: Data Science, IoT, Cybersecurity: Prepare to Data Economy and Industry 4.0
- **DARE project by APEC** (Asia Pacific Economic Cooperation)
  - Continuing cooperation for Asia Pacific region – Workshop 15-16 July 2019, AP
  - Recommended Data Science and Analytics Competences published August 2017 - https://www.apec.org/Press/Features/2017/0620_DSA

EDISON Initiative Online Presence

- EDSF github project - https://github.com/EDISONcommunity/EDSF
  - Component documents CF-DS, DS-BoK, MC-DS, DSPP

- EDISON Community work area and discussions - https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome
- Mailing list - edison-net@list.uva.nl

- EDISON project website - old domain ***edison-project.eu*** expired
  - Legacy information to be moved to https://edisoncommunity.github.io/EDSF/ (later http://edison-project.net/)

Links to EDISON Resources

- EDISON Data Science Framework Release 4 (EDSF)
  https://github.com/EDISONcommunity/EDSF

Component EDSF documents

CF-DS – Data Science Competence Framework
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON01_CF-DS-release4-v11.pdf

DS-BoK – Data Science Body of Knowledge
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON02_DS-BoK-release4-v07.pdf

MC-DS – Data Science Model Curriculum
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON03_MC-DS-release4-v07.pdf

DSPP – Data Science Professional profiles
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON04_DSPP-release4-v08.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
- **Report „Data Science: Lern und Ausbildungsinhulter" (Data Science: Learning and teaching skills)**, December 2019 by the Gesselschaft fur Informatik (GI) workforce "Data Science / Data Literacy" in collaboration with the Ministry of Education and Research (BMBF) and acatech Platform learning system (Germany)
  https://gi.de/fileadmin/GI/Allgemein/PDF/GI_Arbeitspapier_Data-Science_2019-12_01.pdf
- **European Foundational Body of Knowledge for the ICT Profession** (ICT BoK) (CEN TC428 and ICTPE working group)
  https://www.ecompetences.eu/wp-content/uploads/2021/03/210222_CEN_ICT_BoK_all_KUs_DRAFT_v2.pdf