Developing Data Science and Analytics related competences and professional skills

EDSF Components and Practical Use

Yuri Demchenko, EDISON Project
University of Amsterdam

ICDATA19

30 July 2019, Las Vegas

EDISON
building the data science profession

FAIRSFAIR
Fostering Fair Data Practices in Europe

mates

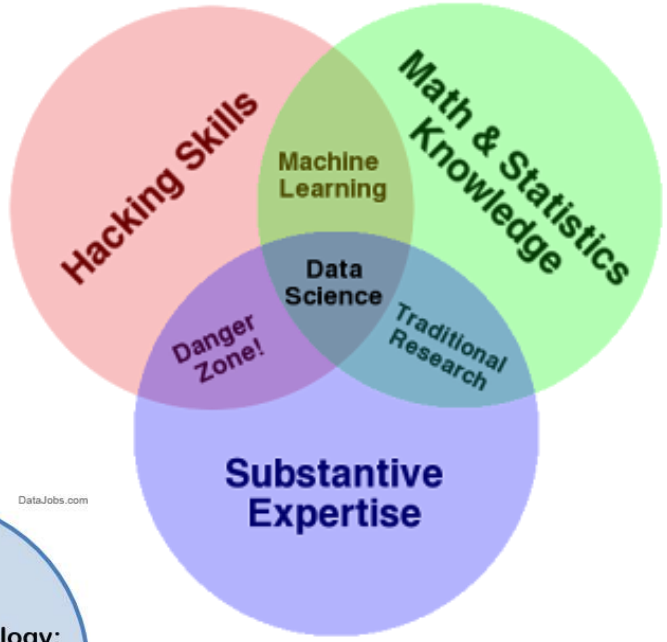
Outline

- Data Scientist definition
- EDSF components
  - CF-DS, DS-BoK, MC-DS, DSPP
  - Data Science Soft and Workplace skills
  - MC-DS and related teaching technologies
- Curriculum design approach
- Competences assessment and Team building
- Data Management and Governance and Research Data Management curricula
- How to become a Data Scientist
  - Online self-learning resources

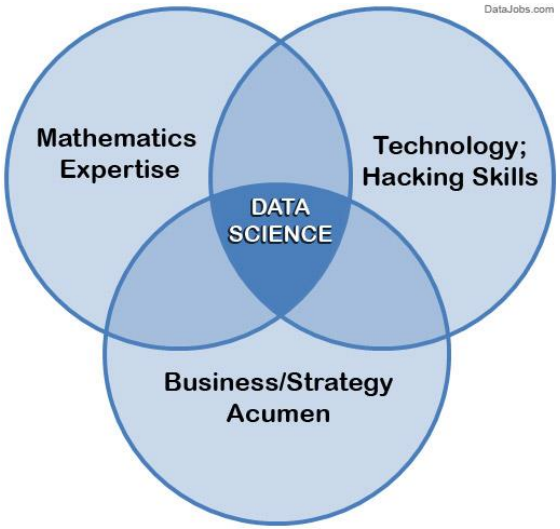
Challenge for Education: Sustainable ICT and Data Skills Development

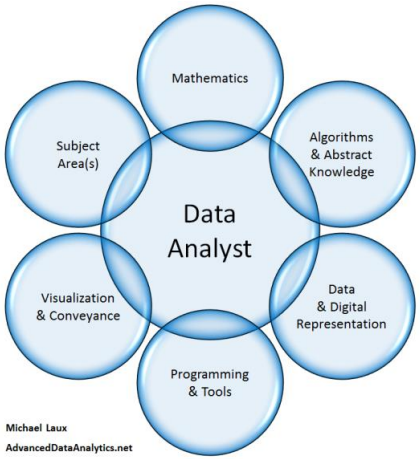
- Educate vs Train
  - Training is a short term solution
  - Education is a basis for sustainable skills development
  - *Importance of workplace or professional attitude skills (not covered in academic curricula)*
- Technology focus changes every 3-4 years
  - Study: 50% of academic curricula are outdated at the time of graduation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

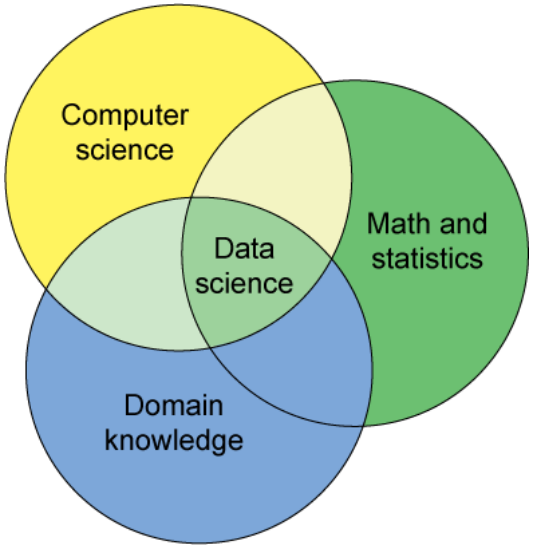
Data Scientist definitions: From Math to Hacking



- Strongly depend on the background of the Data Scientist





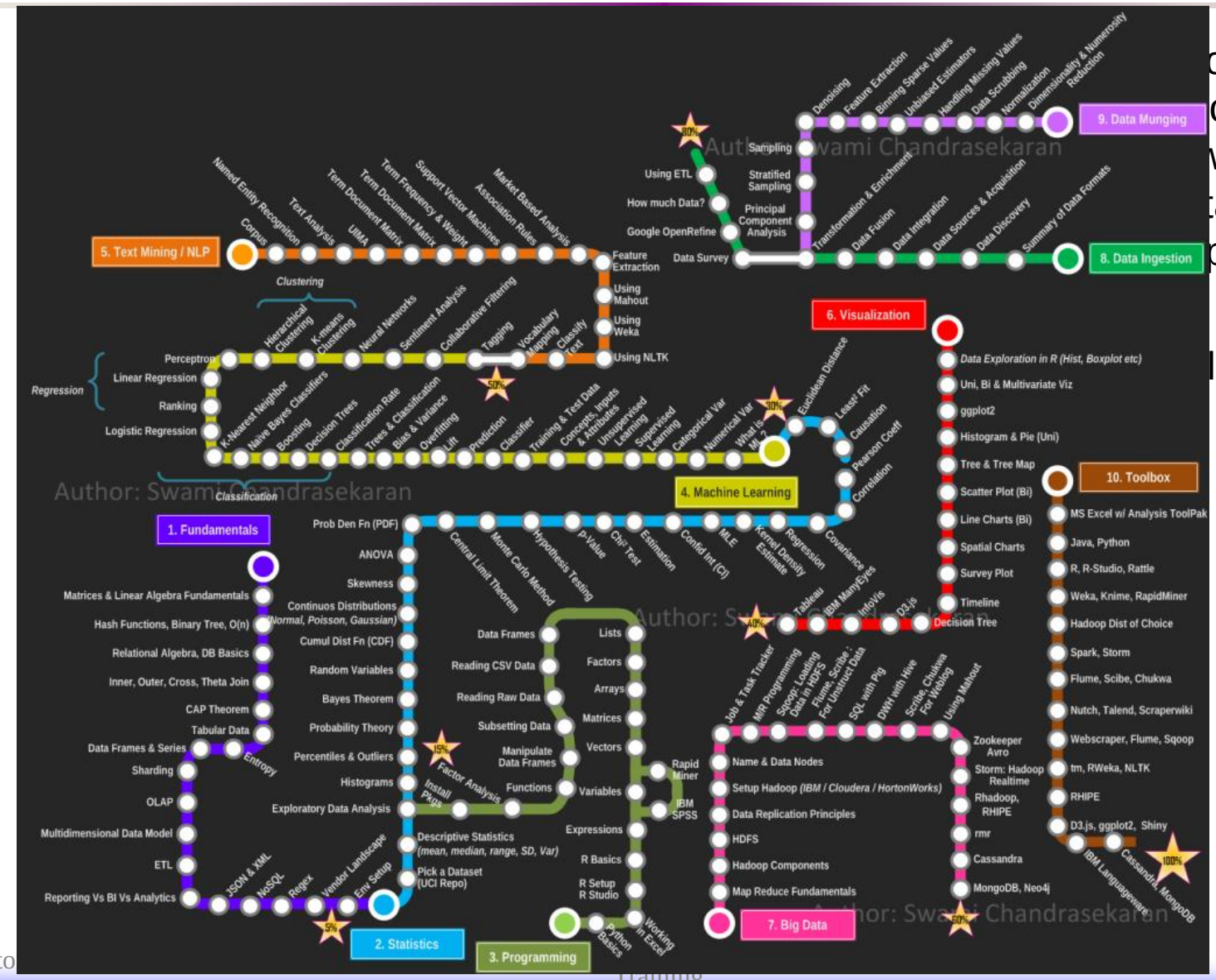


Becoming a Data Scientist by Swami Chandrasekaran (2013)

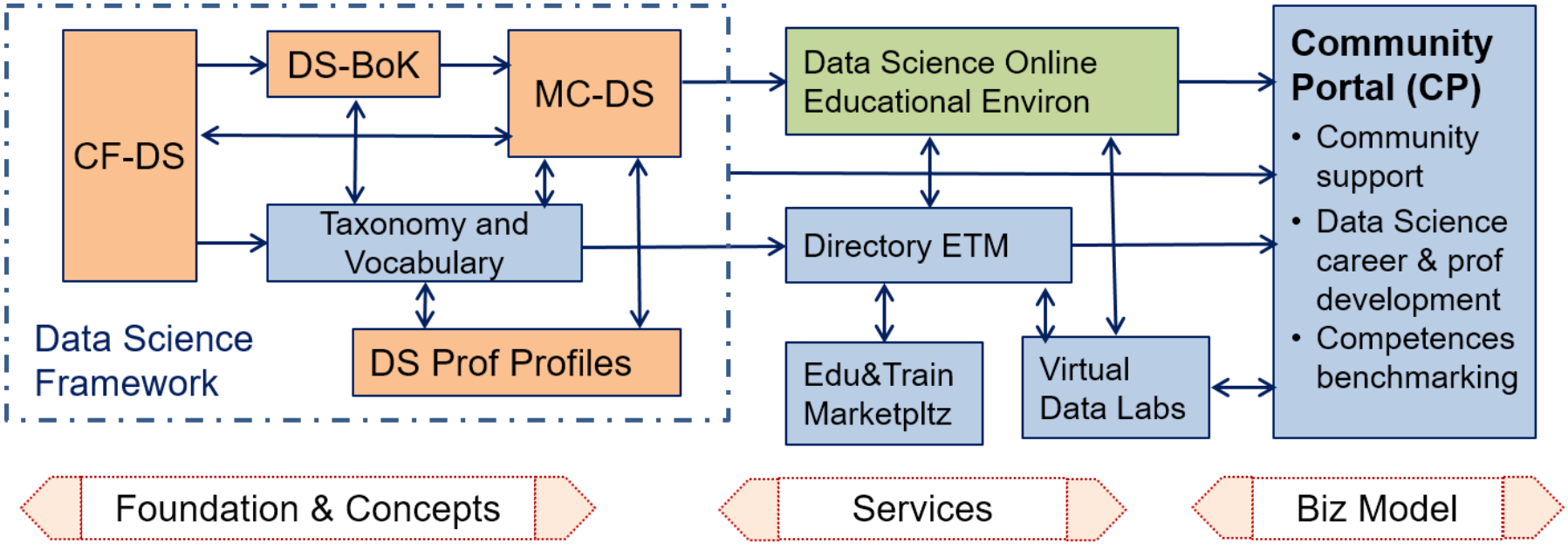
http://nirvacana.com/thoughts/becoming-a-data-scientist/

od and
ctical advice
v to learn
ta Science,
p by step

low the route



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?

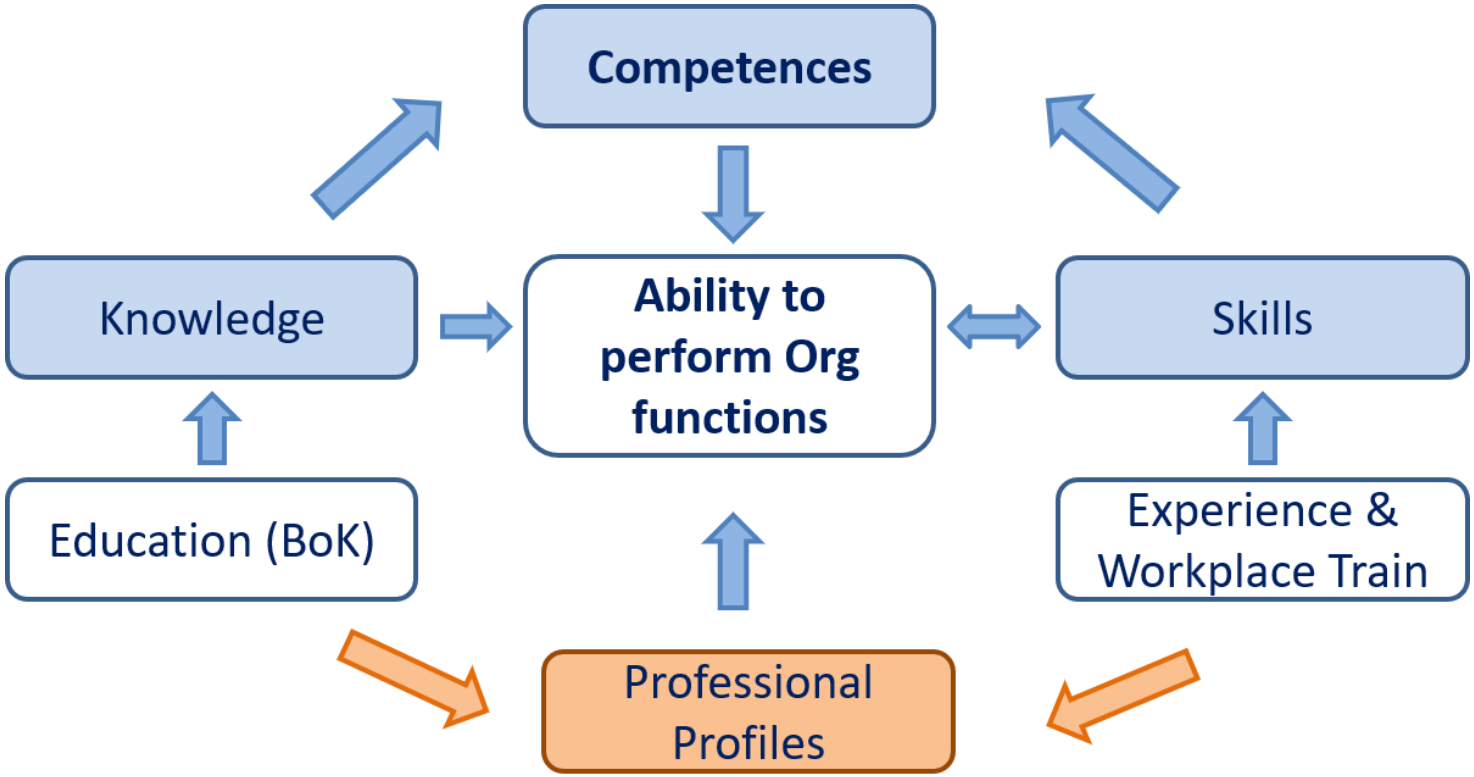
1. Guide researchers in using right methods and tools, latest Data Analytics technologies to extracting value from scientific data
2. Educate and train RI engineers dev to build modern data intensive research infrastructure and understand trends and project for future
3. Develop new data analytics tools 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

EDSF Background: Used Standards and Best Practices

- e-CFv3.0 - European e-Competence Framework for IT
  - Structured by 4 Dimensions and organizational processes
    - Competence Areas: Plan – Build – Run – Enable - Manage
    - Competences: total defined 40 competences
    - Proficiency levels: identified 5 levels linked to professional education levels
    - Skills and Knowledge
- CWA 16458 (2012): European ICT Professional Profiles Family Tree
  - Defines 23 ICT profiles for common ICT jobs
- ESCO (European Skills, Competences, Qualifications and Occupations) framework
  - Standard for European job market since 2016
  - Expected inclusion of the Data Science occupations family – end 2017
- ACM Classification of Computer Science – CCS (2012)
- ACM Computer Science Body of Knowledge (CS-BoK) and ACM and IEEE Computer Science Curricula 2013 (CS2013)
- NIST SP 1500-1 NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions, September 2015

Competences Map to Knowledge and Skills

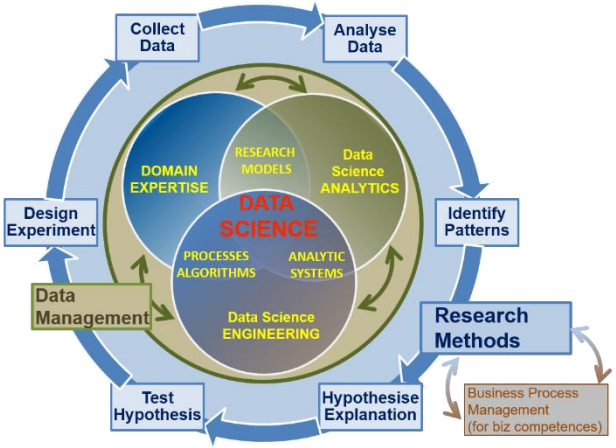
- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results (e-CFv3.0)



Data Scientist definition

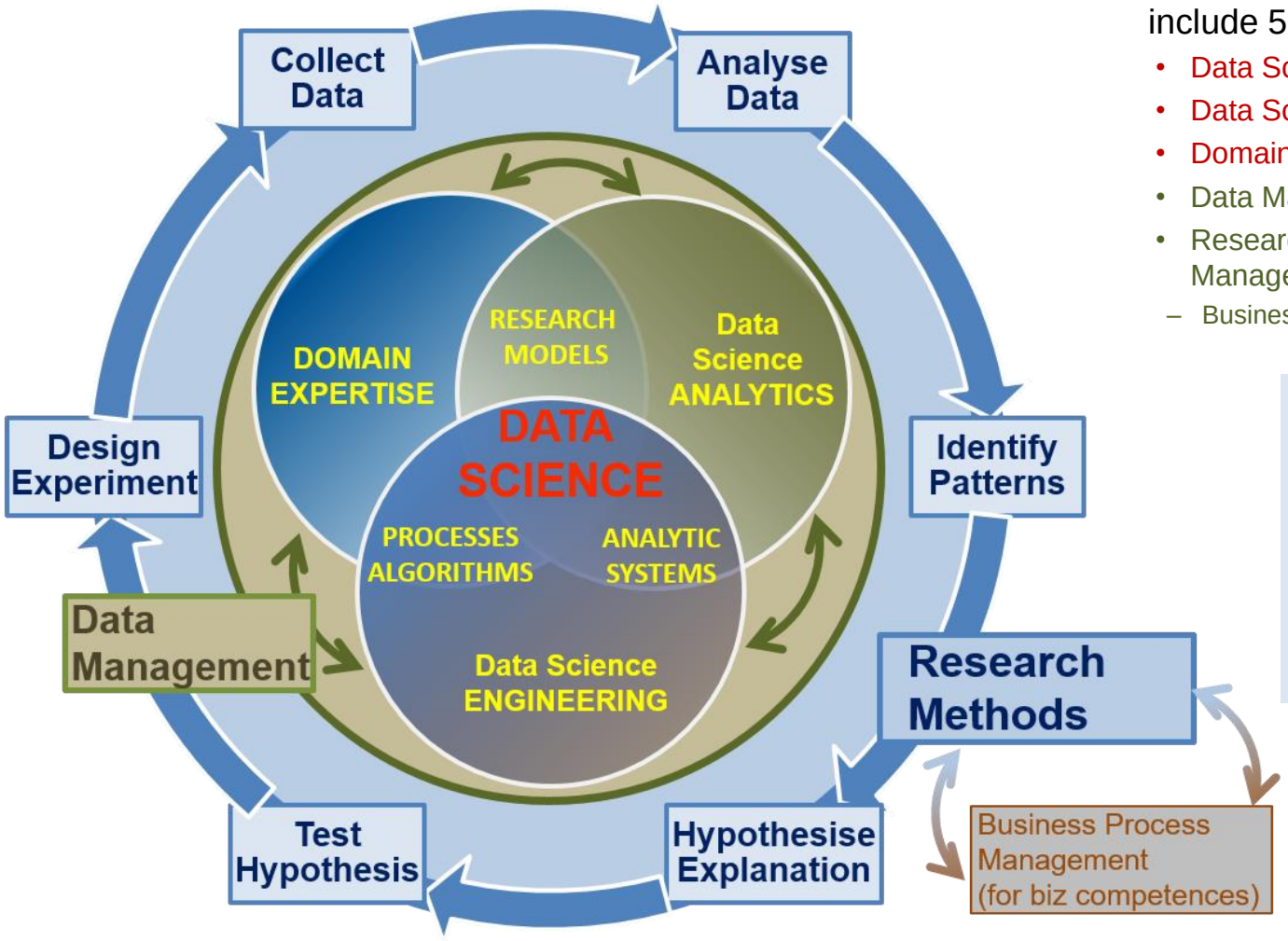
Based on the definitions by NIST SP1500 – 2015, extended by EDISON

- *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)

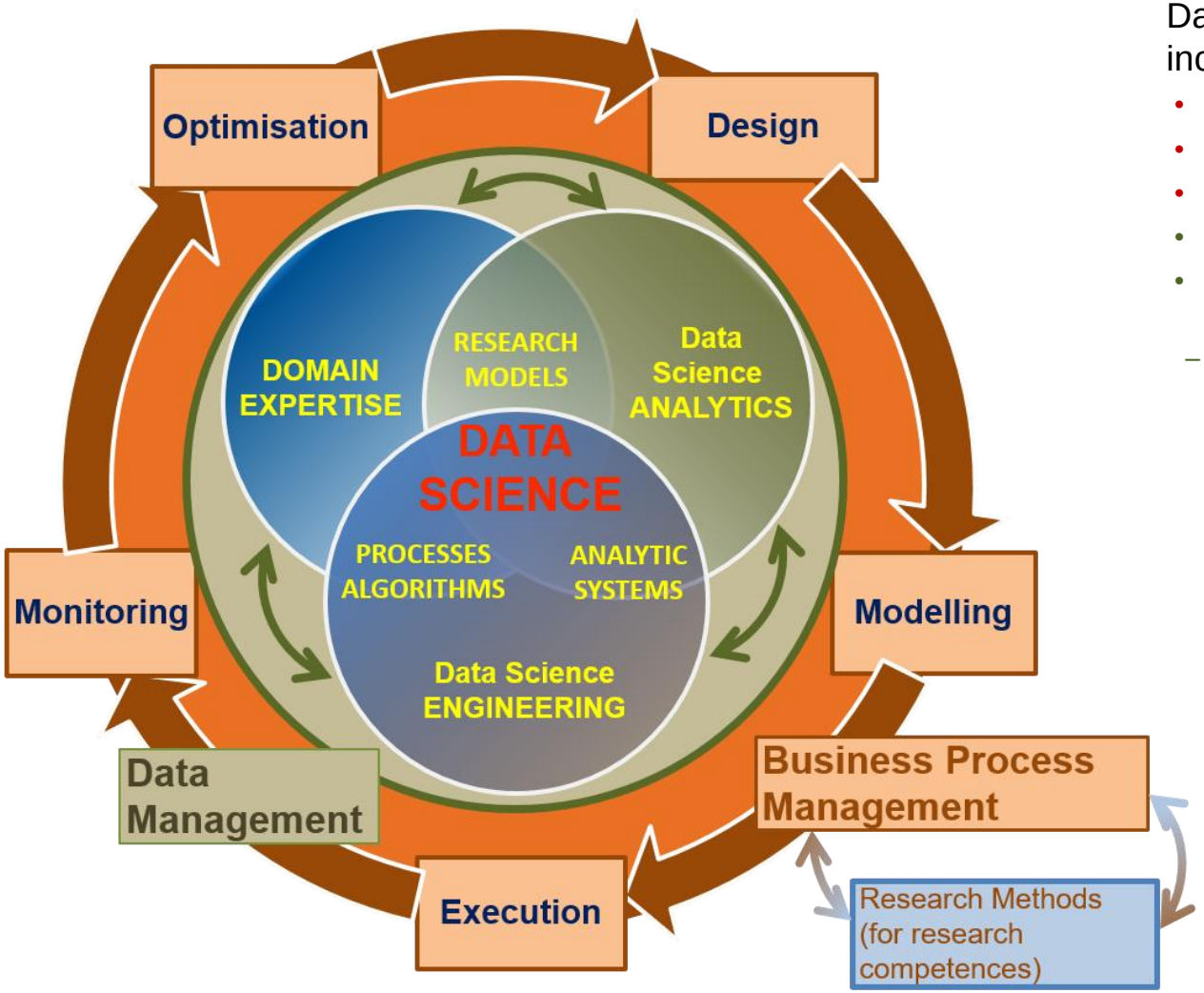
Scientific Methods

- 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

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 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)

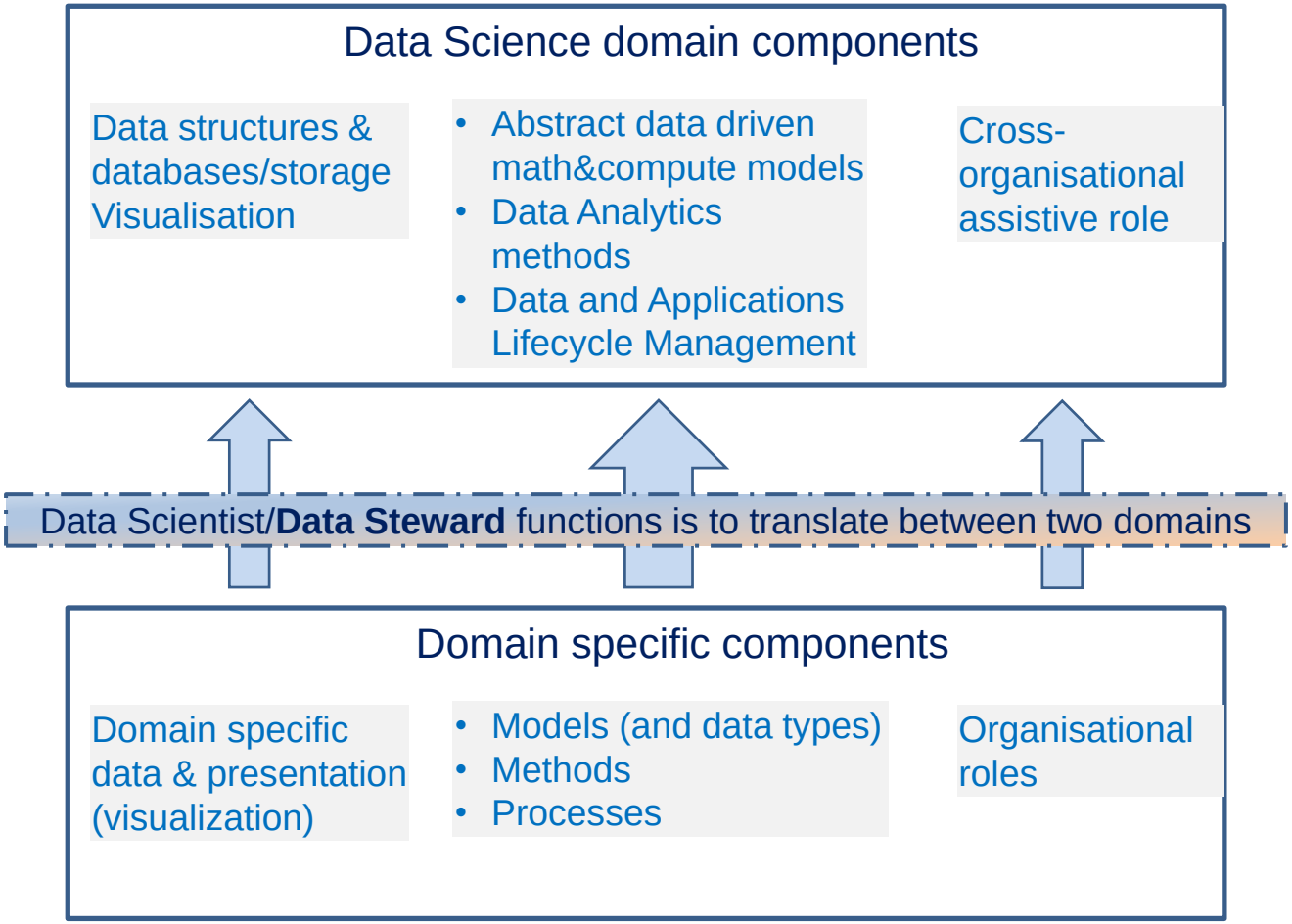
1. **Critical Thinking:** Demonstrating the ability to apply critical thinking skills to solve problems and make effective decisions
2. **Communication:** Understanding and communicating ideas
3. **Collaboration:** Working with other, appreciation of multicultural difference
4. **Creativity and Attitude:** 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

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

CF-DS on the way to standardisation

- Contribution to e-CFv4.0
- Contribution to ICT Professional profiles

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

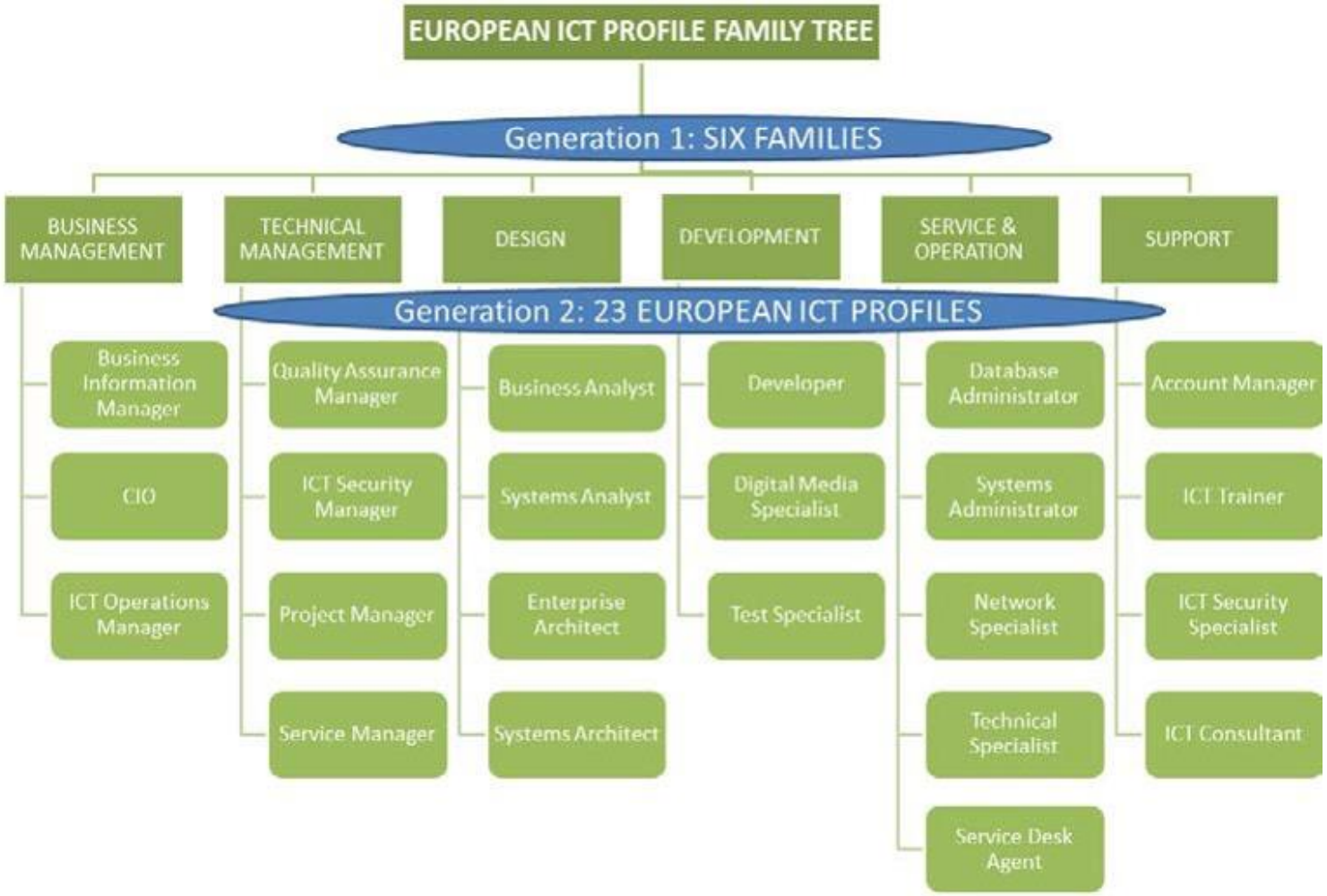
Dimension 1 Competence Group	DSDA	Data Science Analytics
Dimension 2 Competence	DSDA01	Effectively use variety of data analytics techniques, such as Machine Learning (including supervised, unsupervised, semi-supervised learning), Data Mining, Prescriptive and Predictive Analytics, for complex data analysis 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	

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 ID	Knowledge unit definition
Knowledge	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

Defining Data Science Professional Profiles

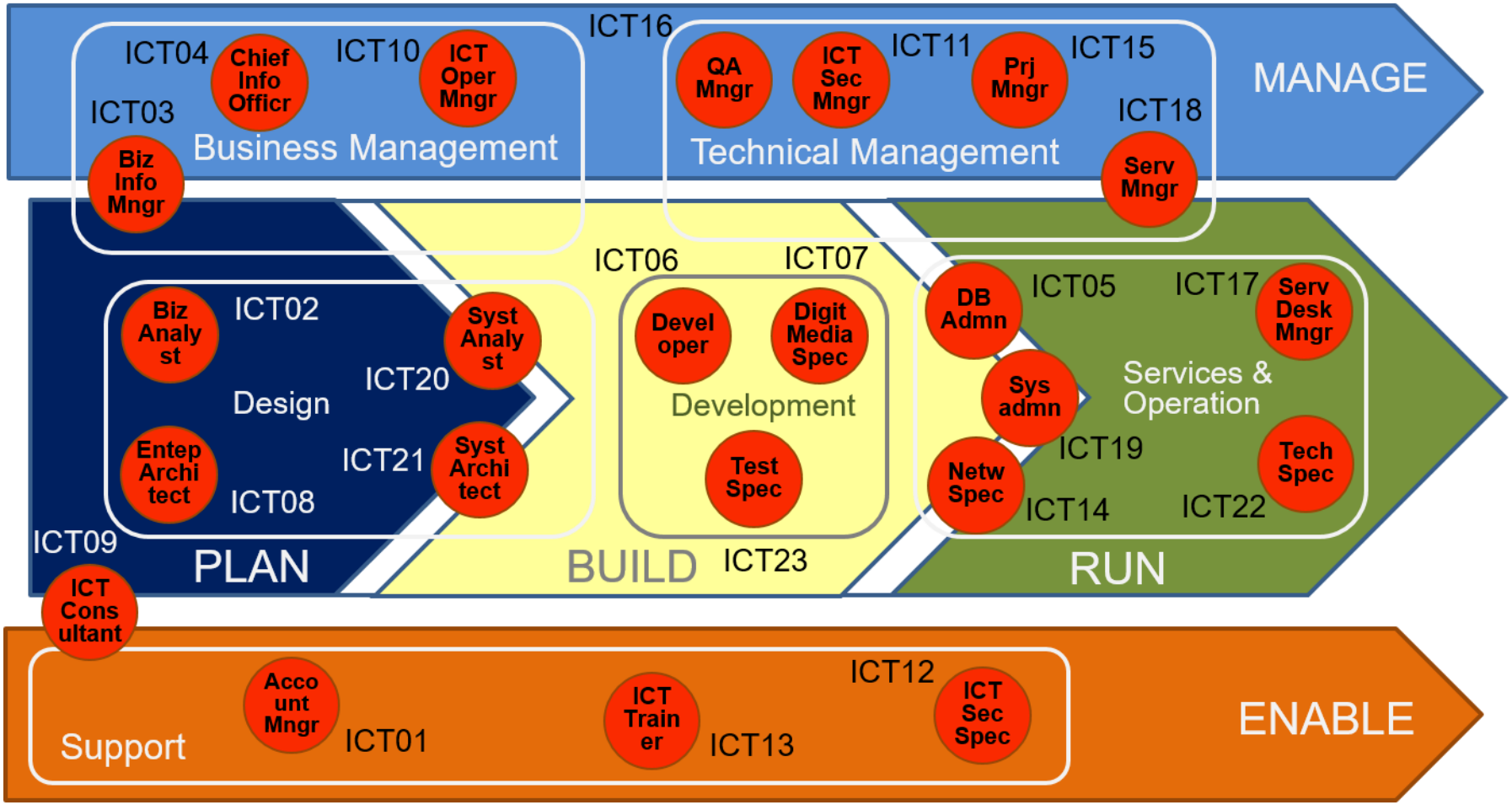
- CWA 16458 (2012): European ICT Professional Profiles
- ESCO (2017): European, Skills, Competences, Occupations

CWA 16458 (2012): European ICT Professional Profiles

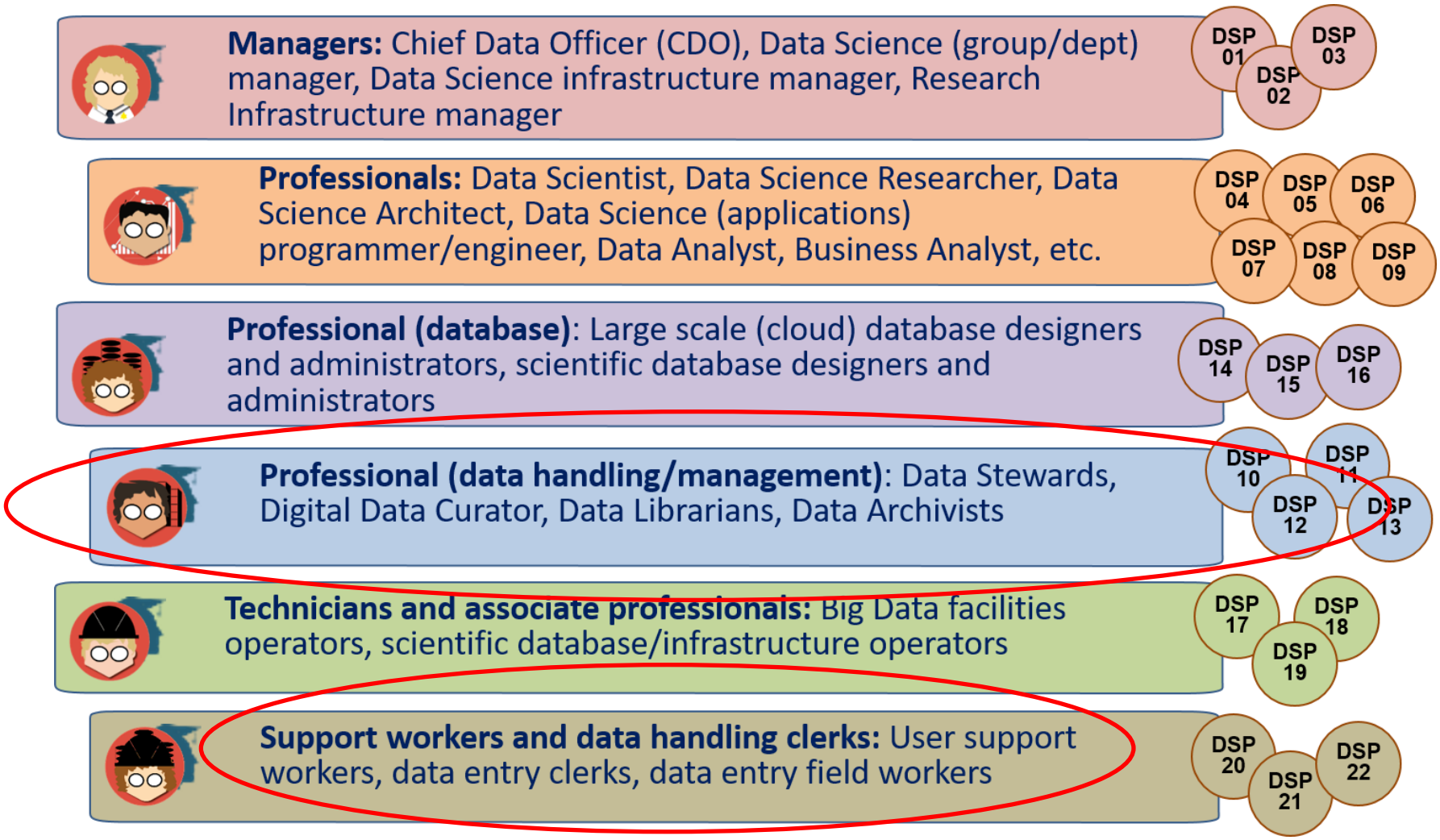


- The CWA defines 23 main ICT profiles the most widely used by organisations

CWA Professional Profiles and Organisational Workflow

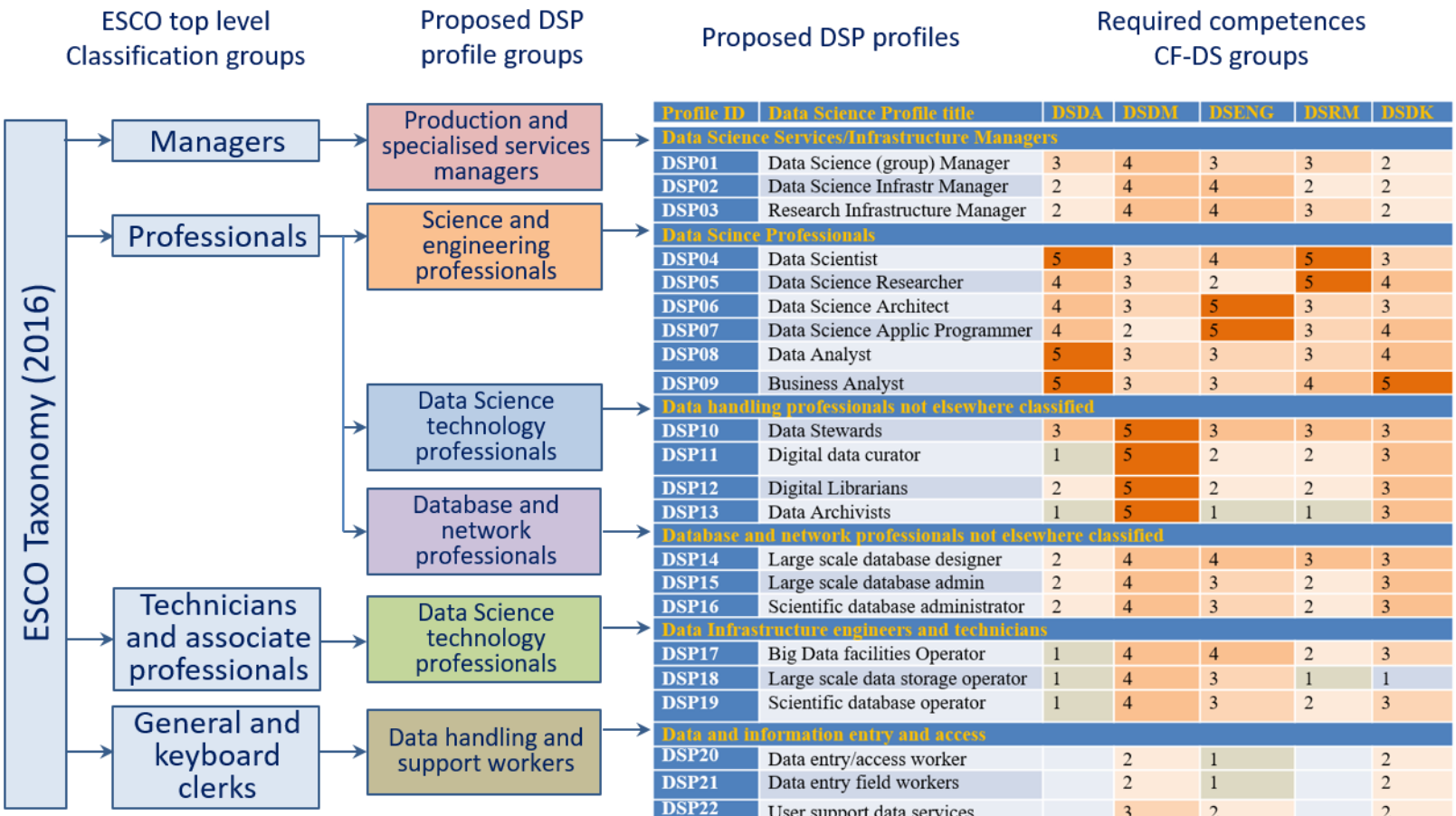


Data Science Professions Family



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

DSP Profiles mapping to ESCO Taxonomy High Level Groups

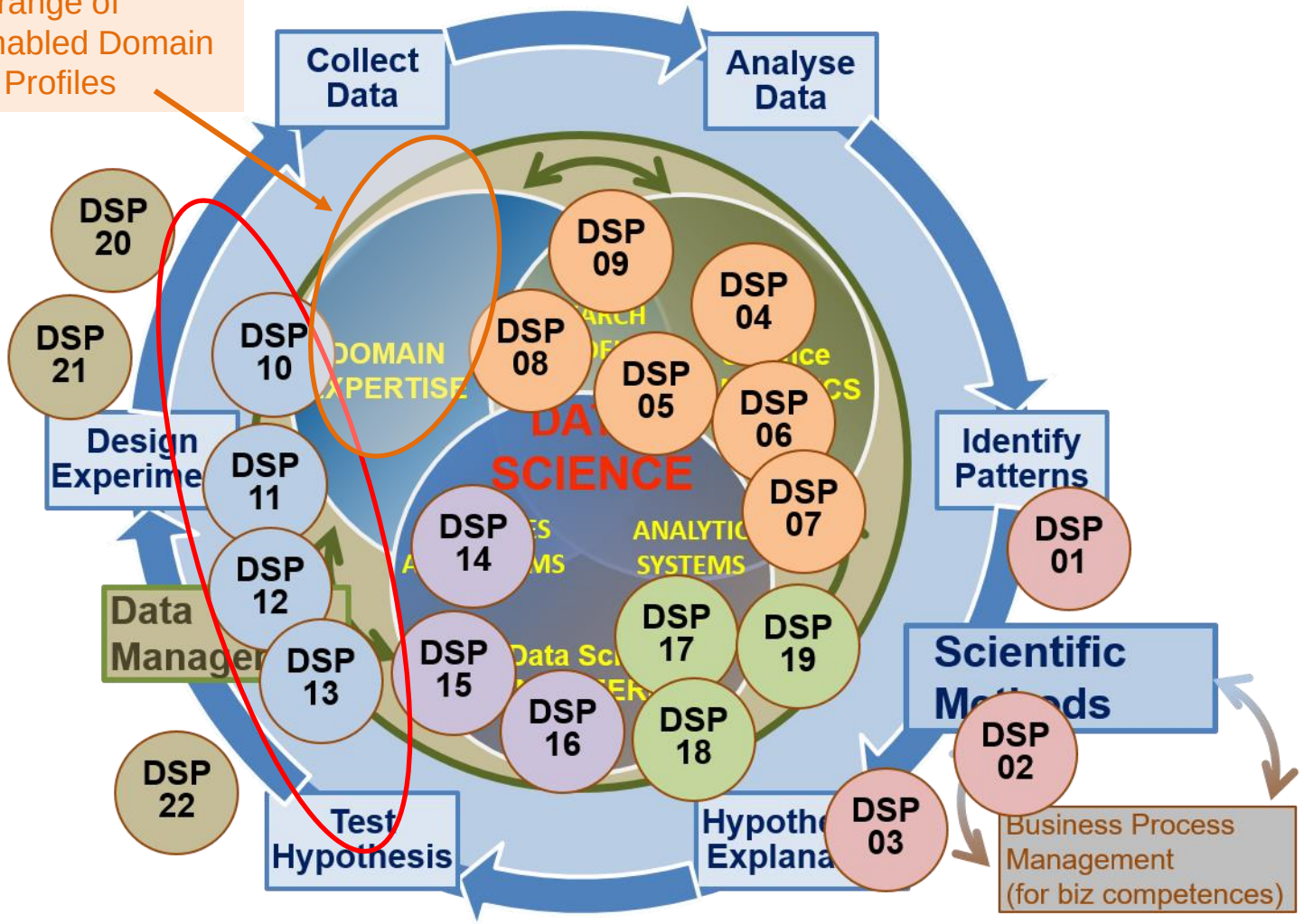


Profile ID	Data Science Profile title	DSDA	DSDM	DSENG	DSRM	DSDK
Data Science Services/Infrastructure Managers						
DSP01	Data Science (group) Manager	3	4	3	3	2
DSP02	Data Science Infrastr Manager	2	4	4	2	2
DSP03	Research Infrastructure Manager	2	4	4	3	2
Data Scince Professionals						
DSP04	Data Scientist	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

Whole range of DSA enabled Domain related Profiles



Template DS Professional Profile Definition (compliant with CWA)

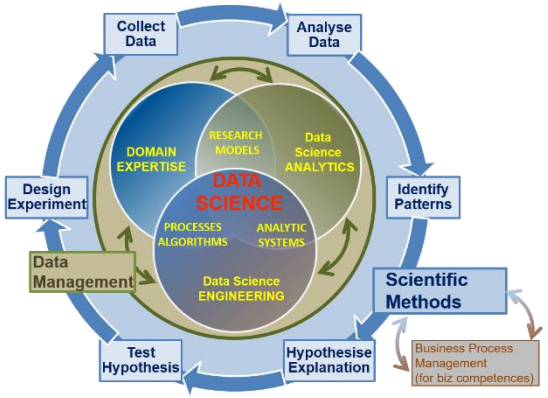
Profile title	**Gives a commonly used name to a profile. TEMPLATE**		
Summary statement	**Indicates the main purpose of the profile.** **The purpose is to present to stakeholders and users a brief, concise understanding of the specified ICT Profile. It should be understandable by ICT professionals, ICT 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 Outcomes 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

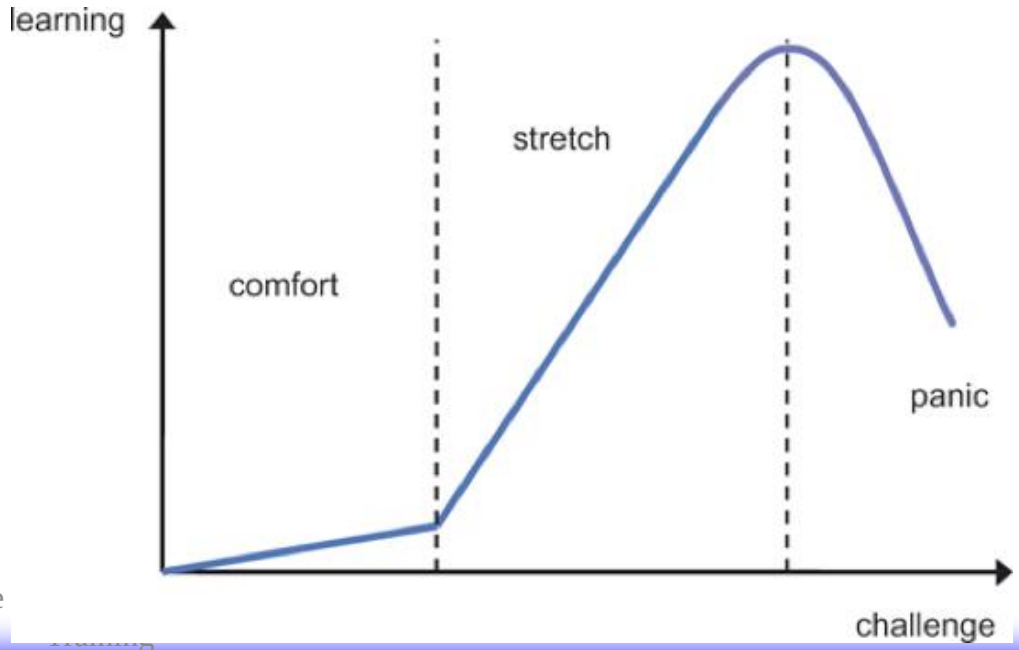
Data Science Model Curriculum (MC-DS)

Data Science Model Curriculum includes

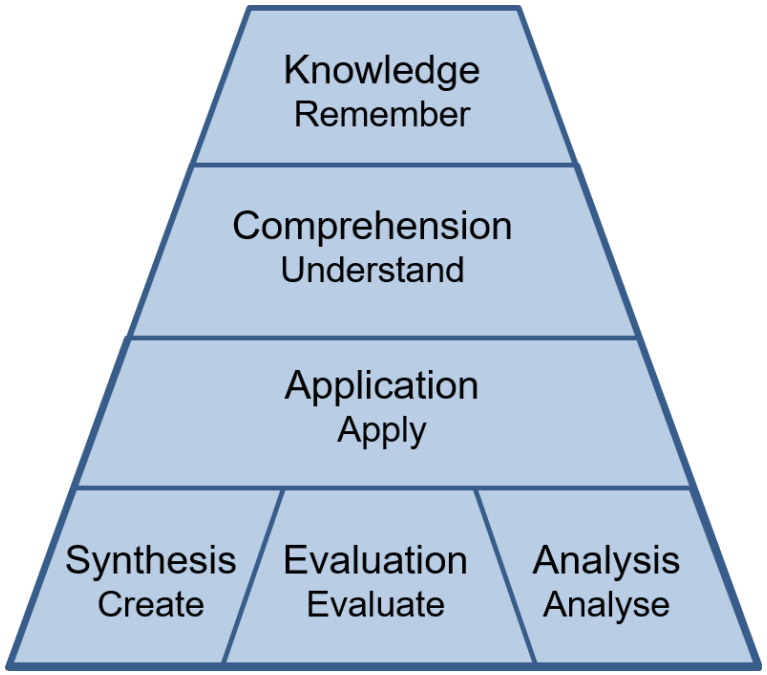
- Learning Outcomes (LO) definition based on CF-DS
  - LOs are defined for CF-DS competence groups and for all enumerated competences
  - Knowledge levels: Familiarity, Usage, Assessment (based in Bloom's Taxonomy)
- LOs mapping to Learning Units (LU)
  - LUs are based on CCS(2012) and universities best practices
  - Data Science university programmes and courses inventory (interactive) http://edison-project.net/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

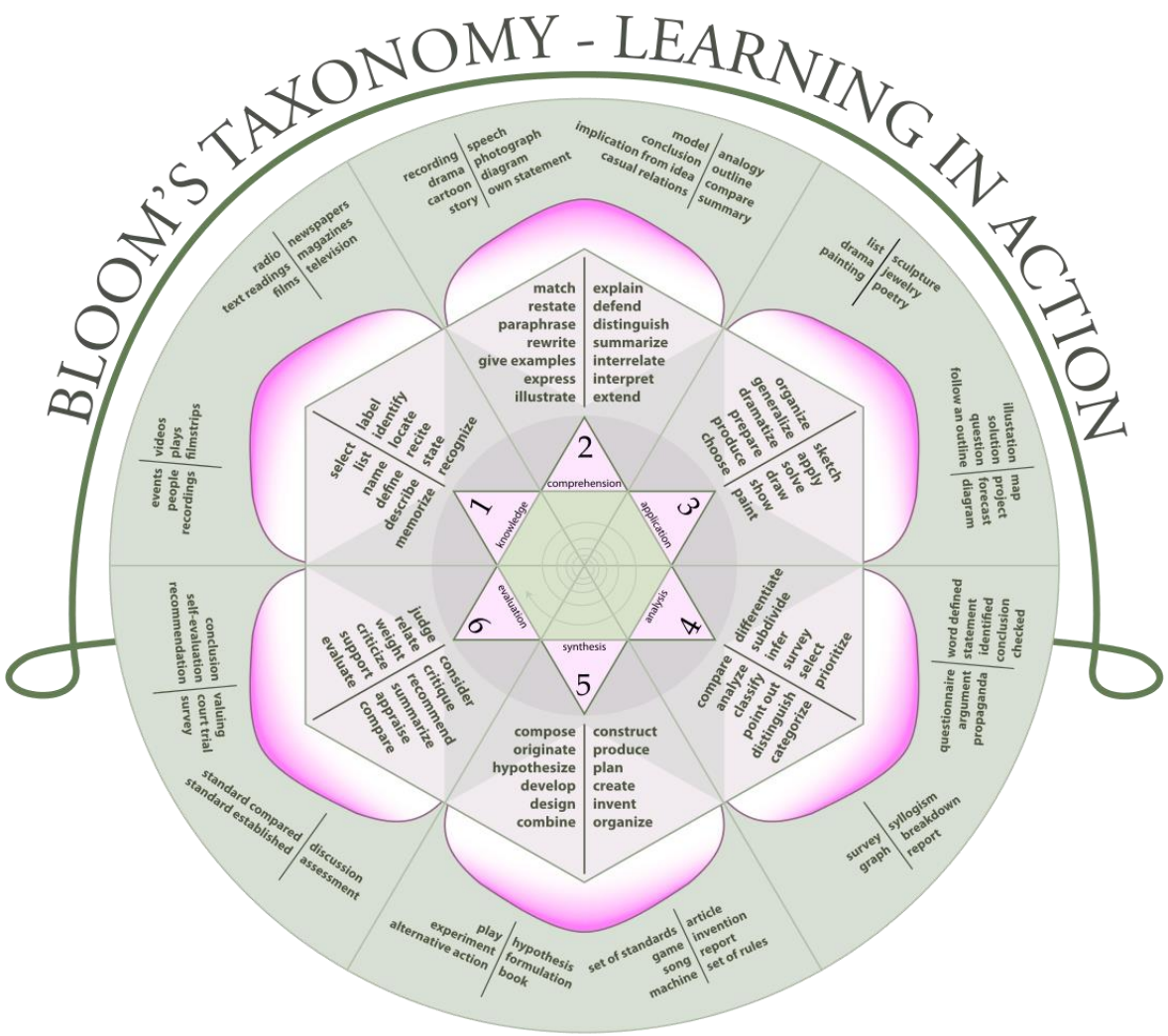


Bloom's Taxonomy and Knowledge Levels for MC-DS



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?**

EDSF-MC 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

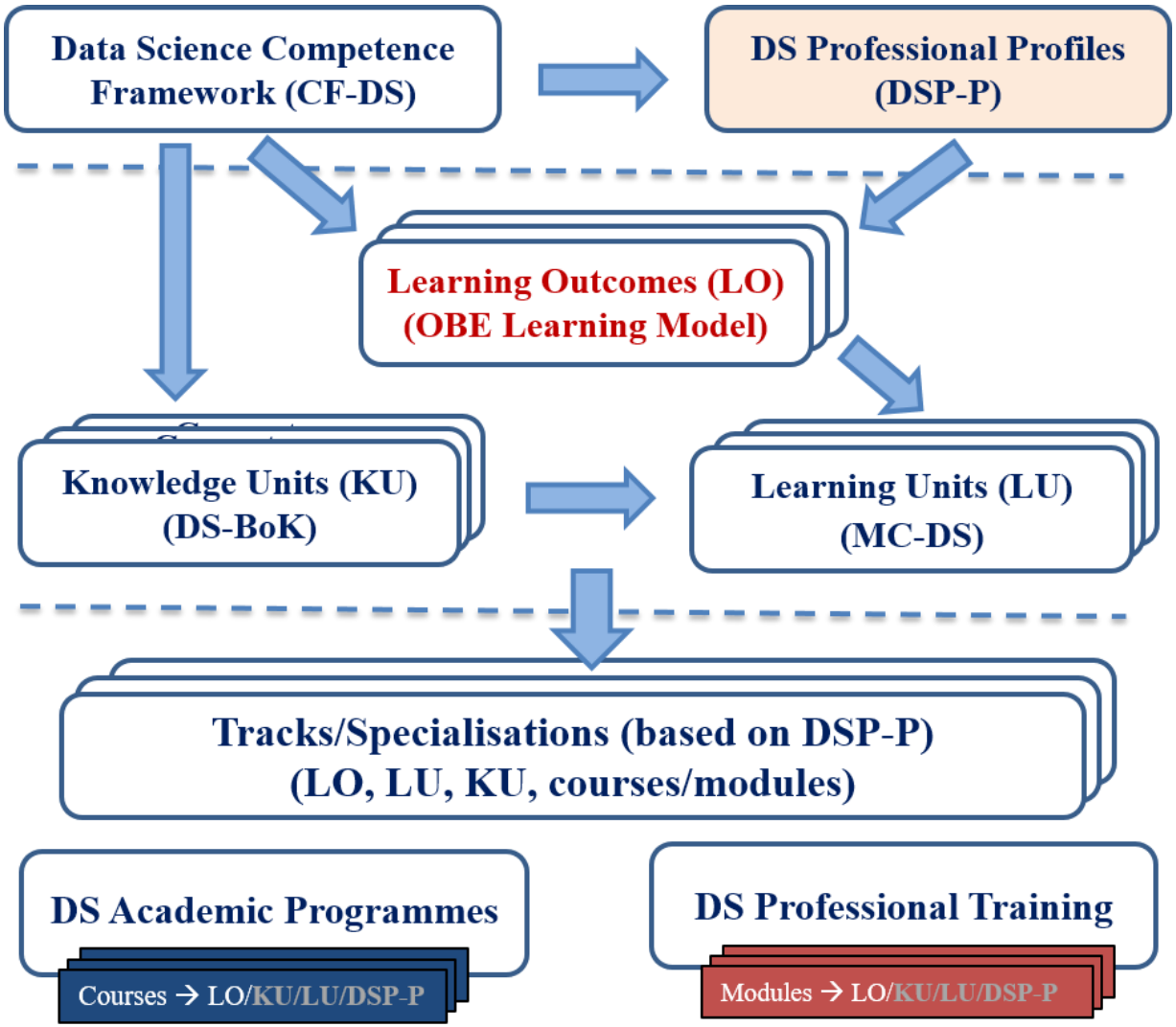
(13) Data Management Plan

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

(15) Responsible data use

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

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



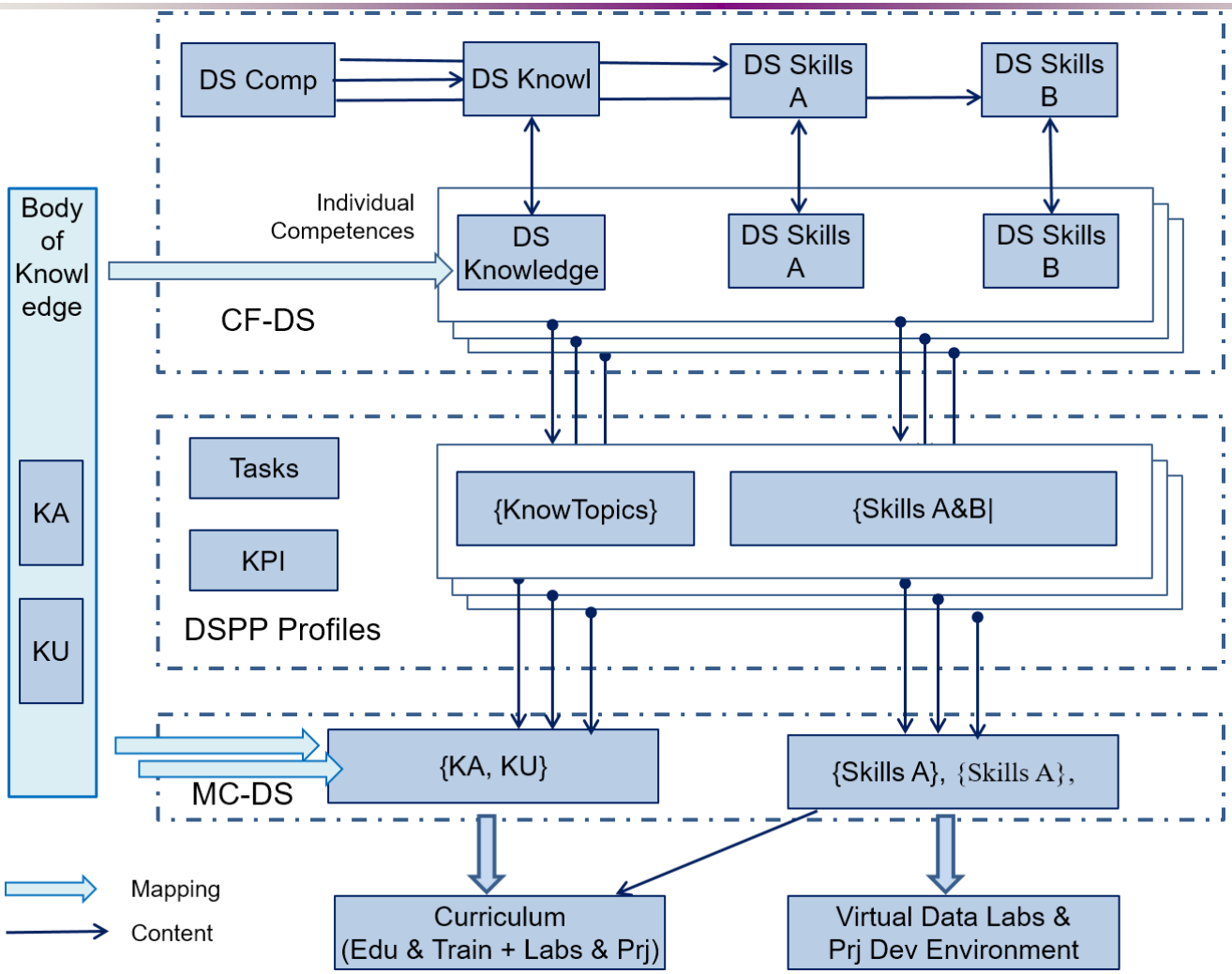
From Competences and DSP Profiles

to Learning Outcomes (LO) and

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

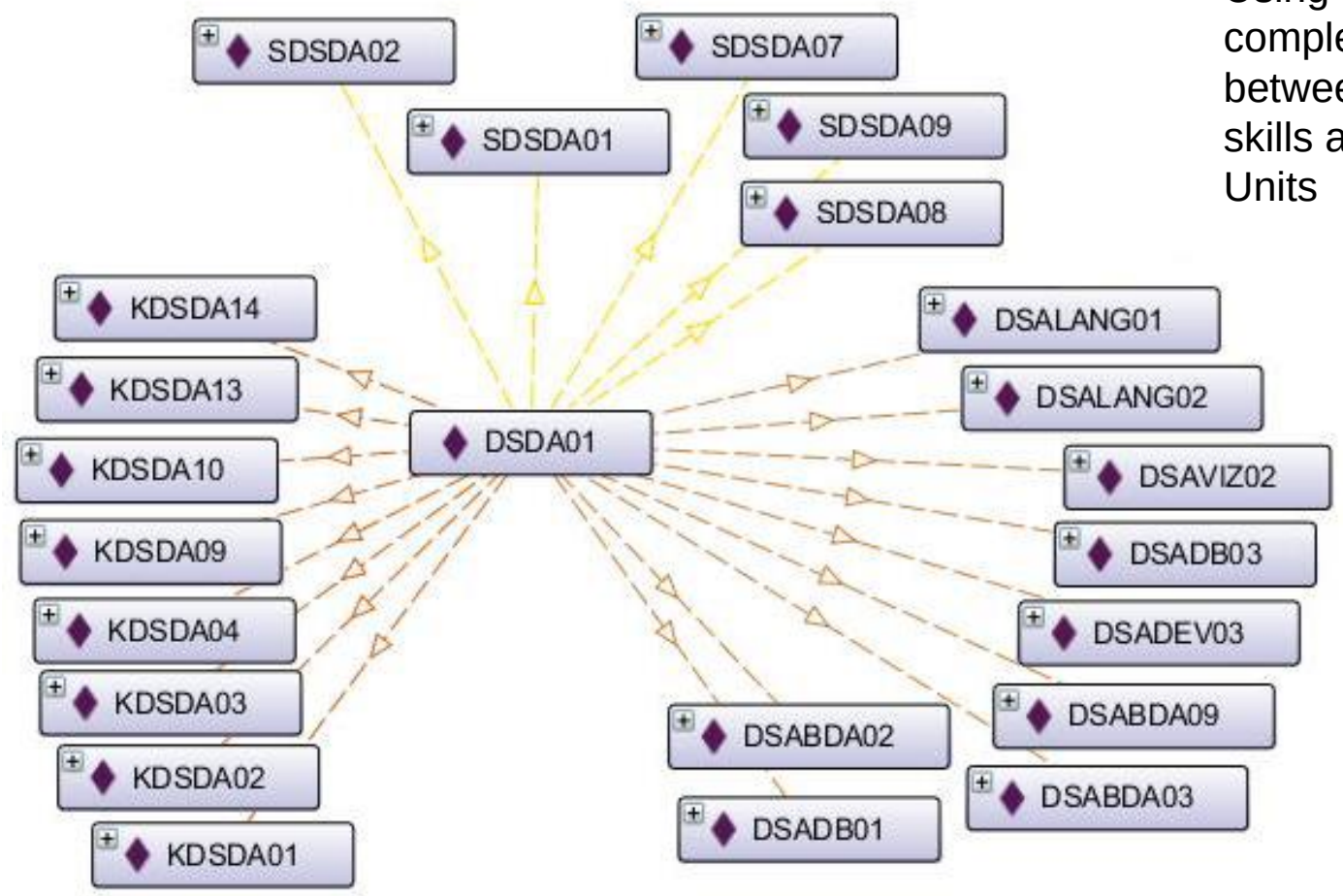
- EDSF allow for customized educational courses and training modules design

EDSF Data Model and API



- EDSF API provides access to all EDSF functionality

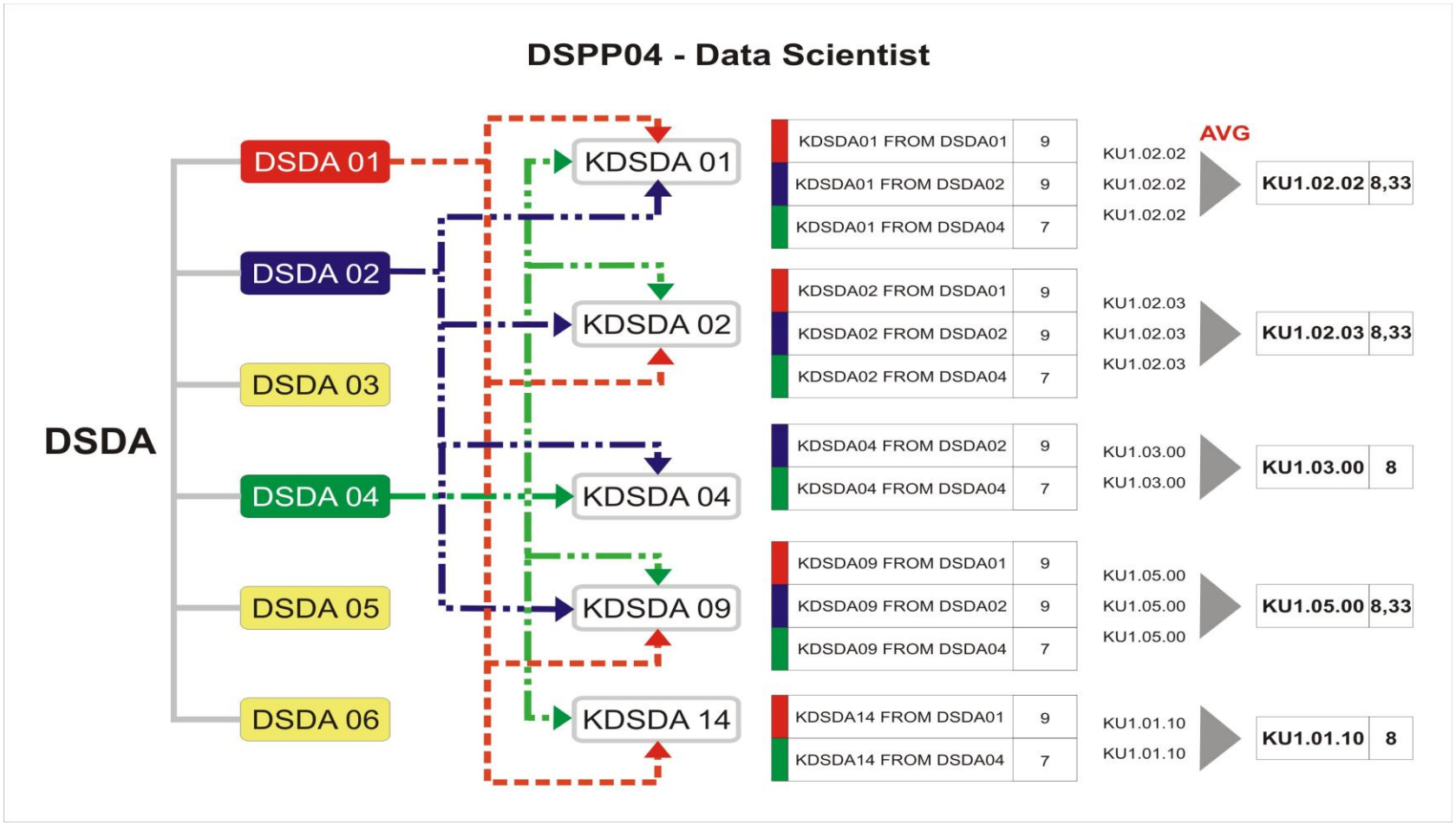
Example DSDA01 Competence and its properties



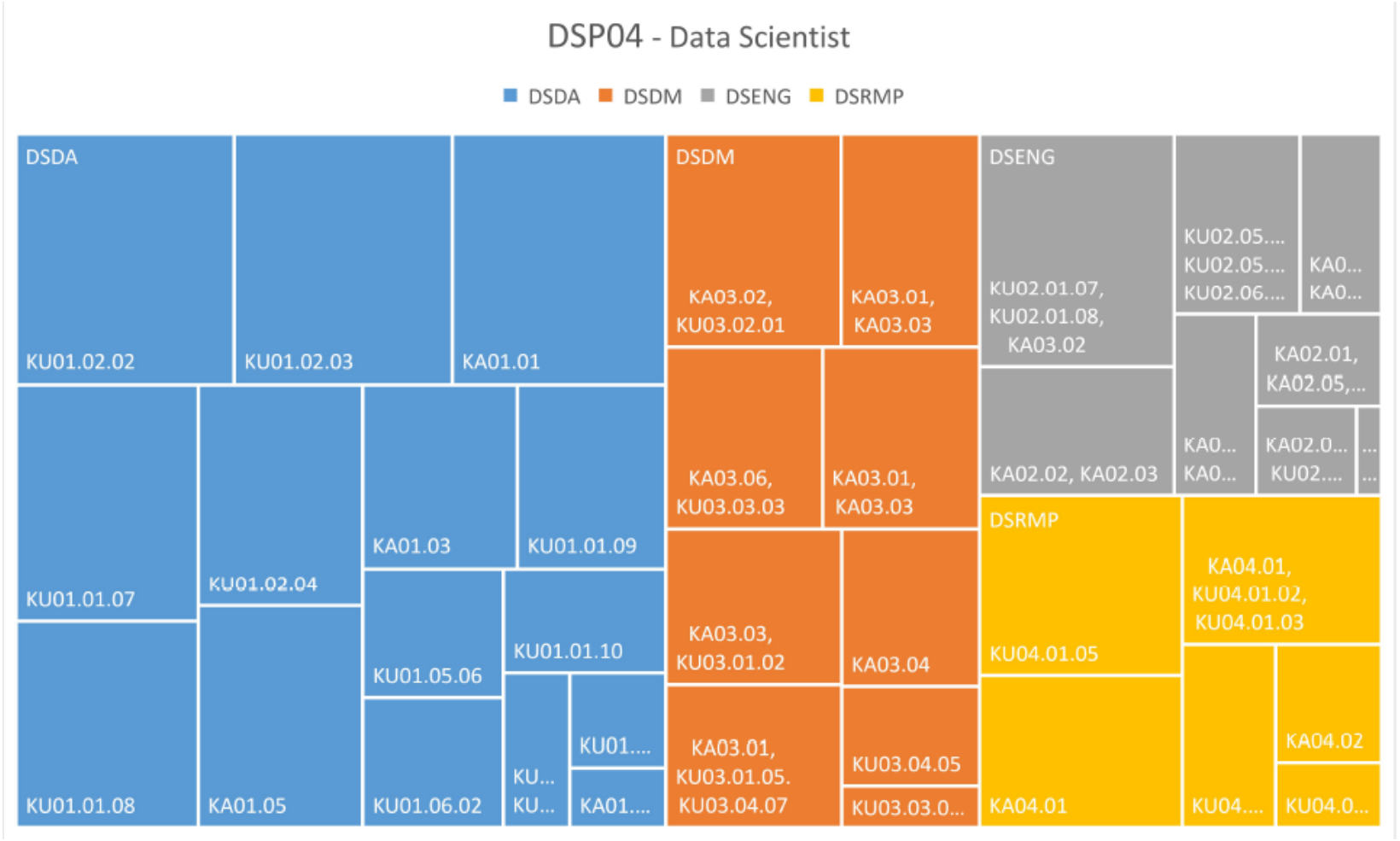
- Using ontology to manage all complexity of relations between Knowledge topics, skills and BoK Knowledge Units

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

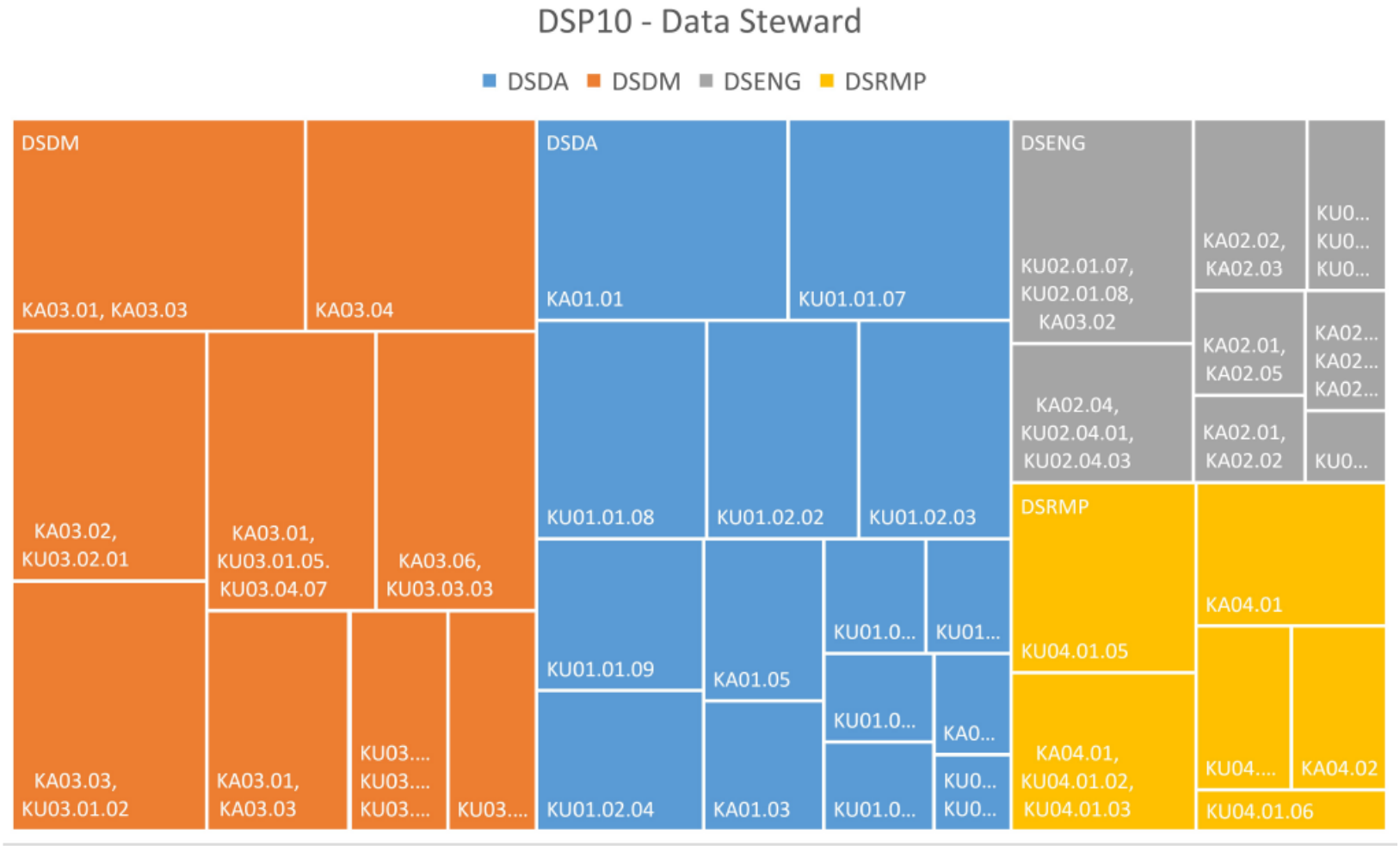
Extracting required Knowledge Units from EDSF ontology for DSPP04 – Data Scientist



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

Hybrid Data Science Education Environment (DSEE)

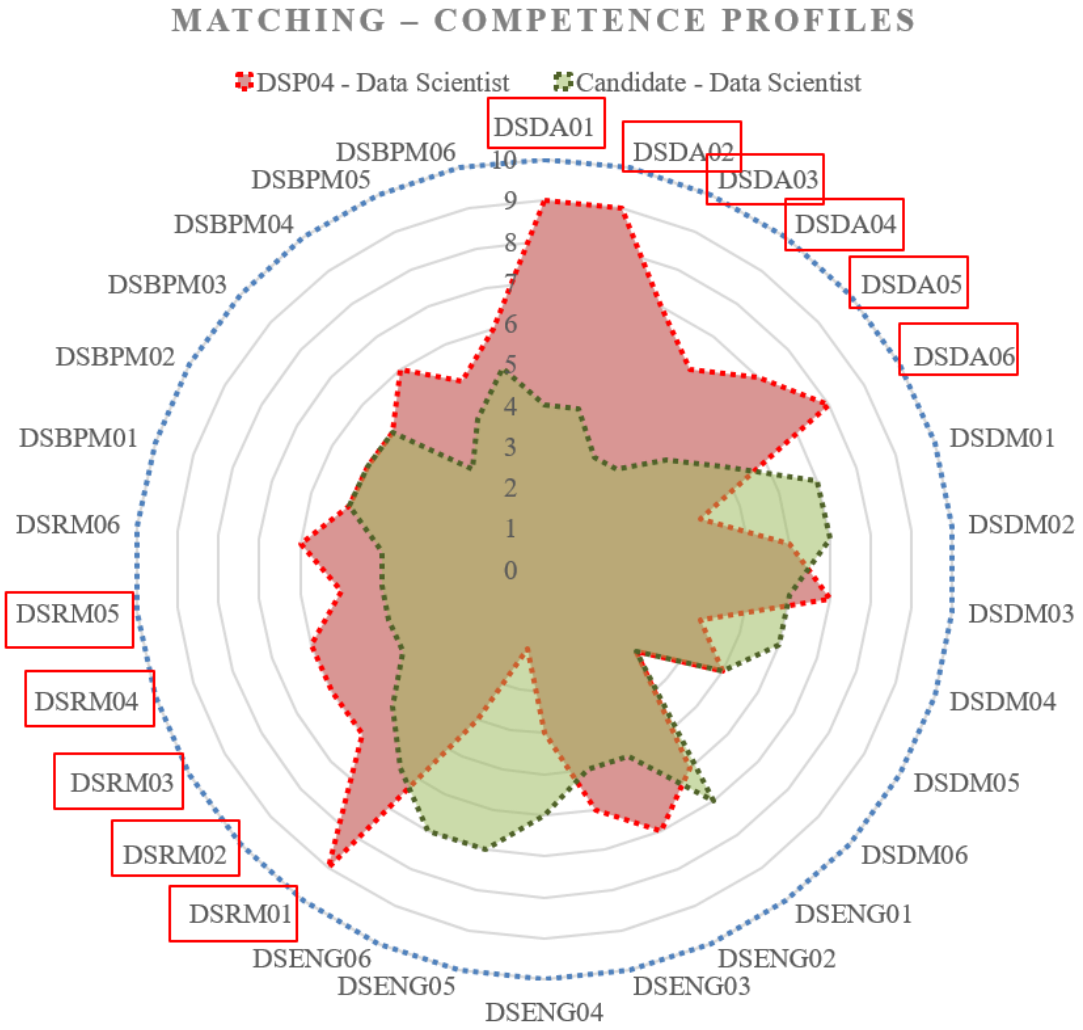
Hybrid DSEE and VDLabs extends regular compute and storage resources with cloud based

- Microsoft Azure Data Lakes Analytics, Power BI, HDInsight Hadoop as a Service, others
- AWS Elastic MapReduce (EMR), QuickSight, Kinesis and wide collection of open datasets
- IBM Data Science Experience, Data Labs, Watson Analytics
- Google Cloud Platform (GCP) with powerful ML and text analysis tools

Competences assessment and Team building

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

Individual Competences Benchmarking

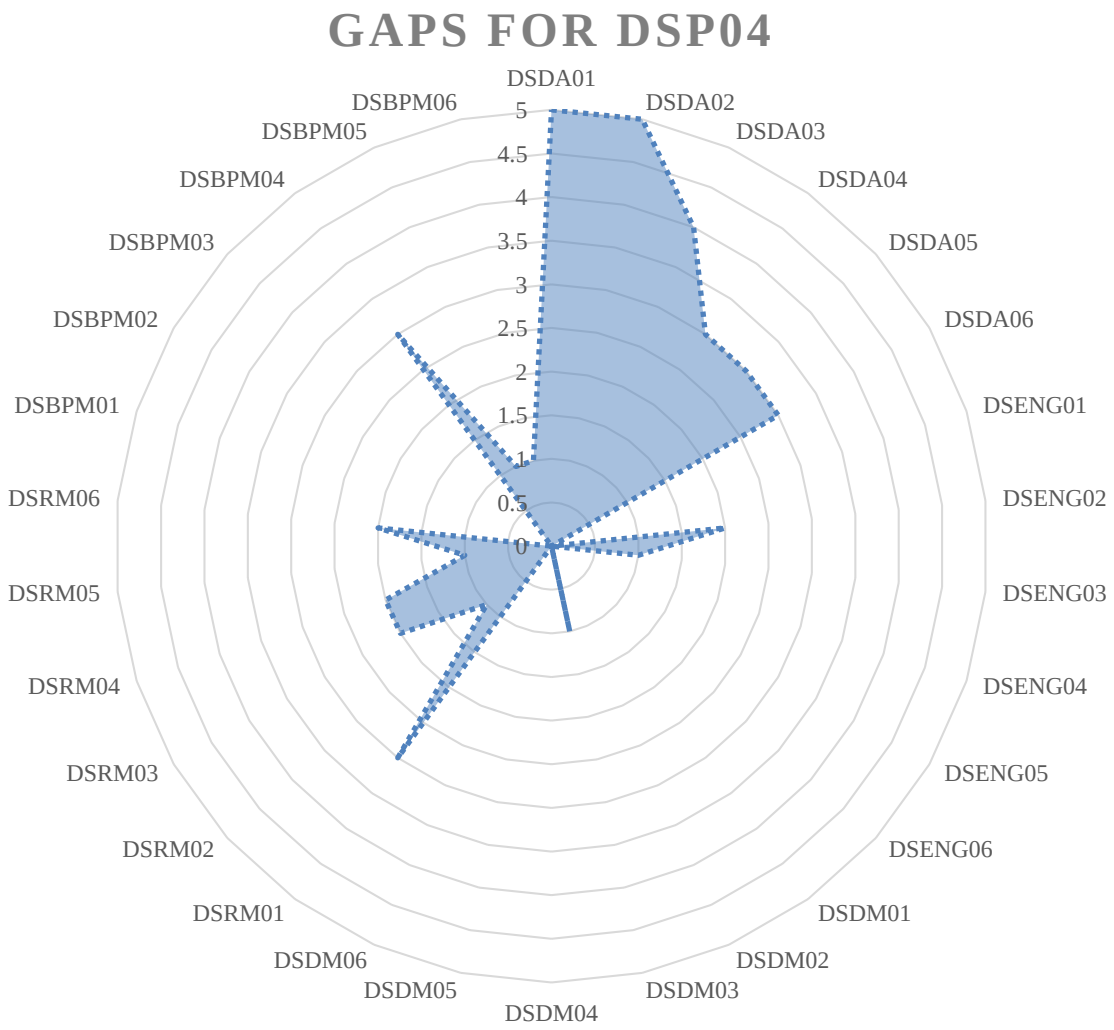


Individual Education/Training Path based on Competence benchmarking

- 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

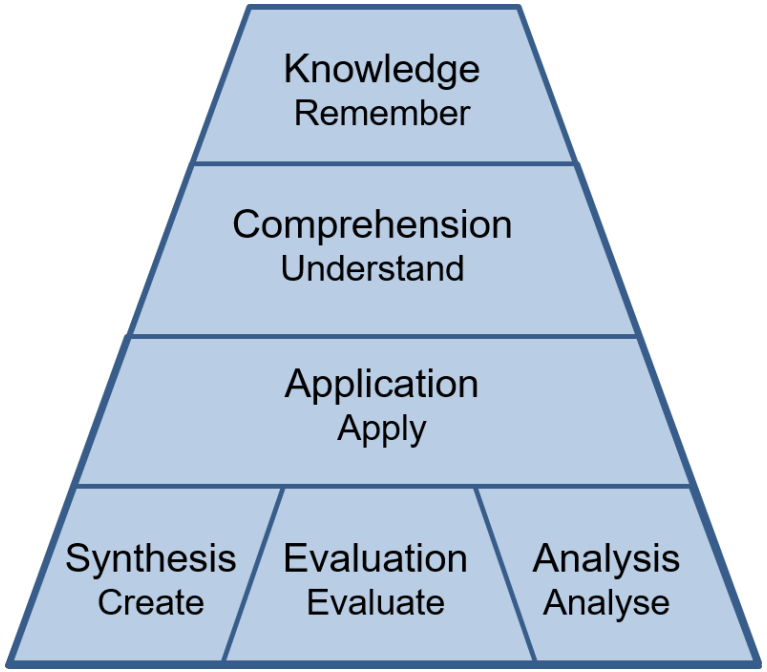
Mastery levels defined using workplace terminology that can be easy mapped to mastery levels defined in MC-DS:

- 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

Bloom's Taxonomy and Knowledge Levels for MC-DS and Curriculum structure limitation

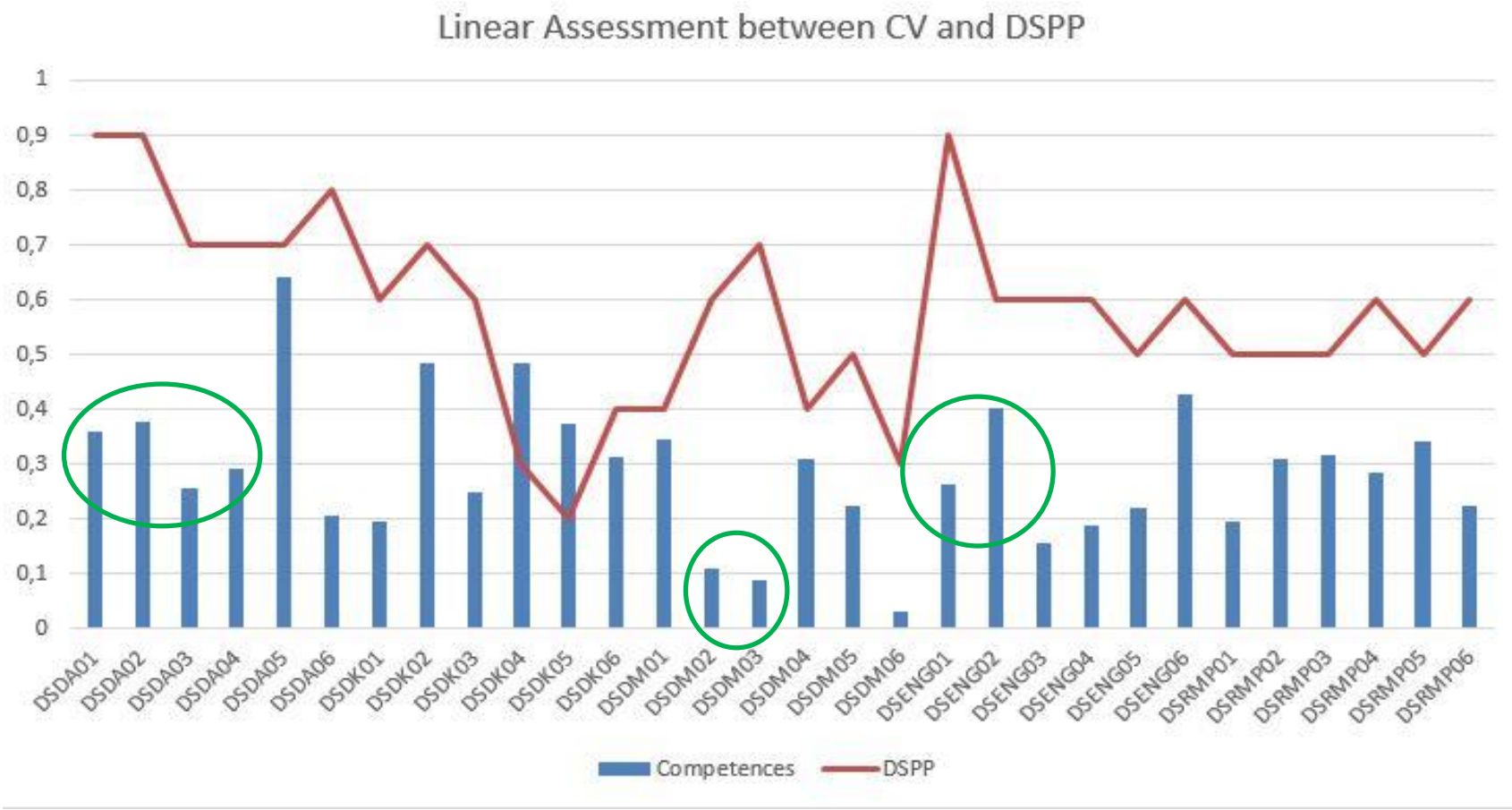


1) Knowledge, remember – min 4 weeks
2) Comprehension, Understanding – 2 months (based on (1))
3) Application, Development – 2-6 months

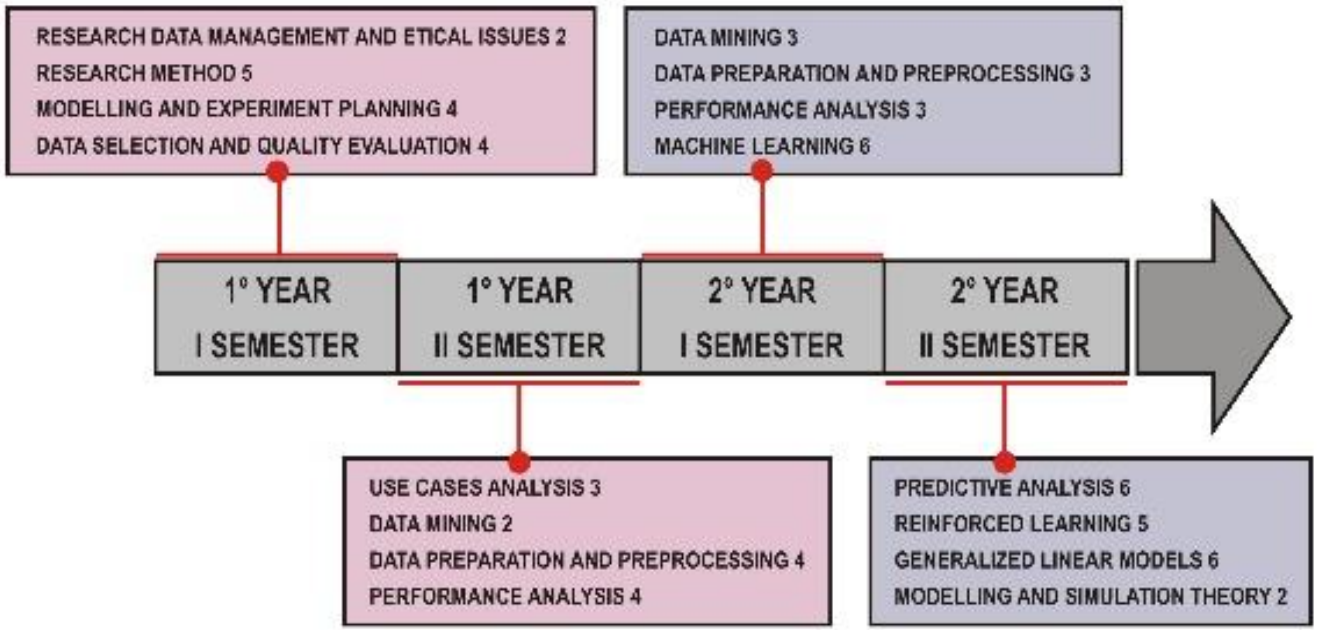
More than 2 months

4) Analysis
5) Evaluation
6) Synthesis

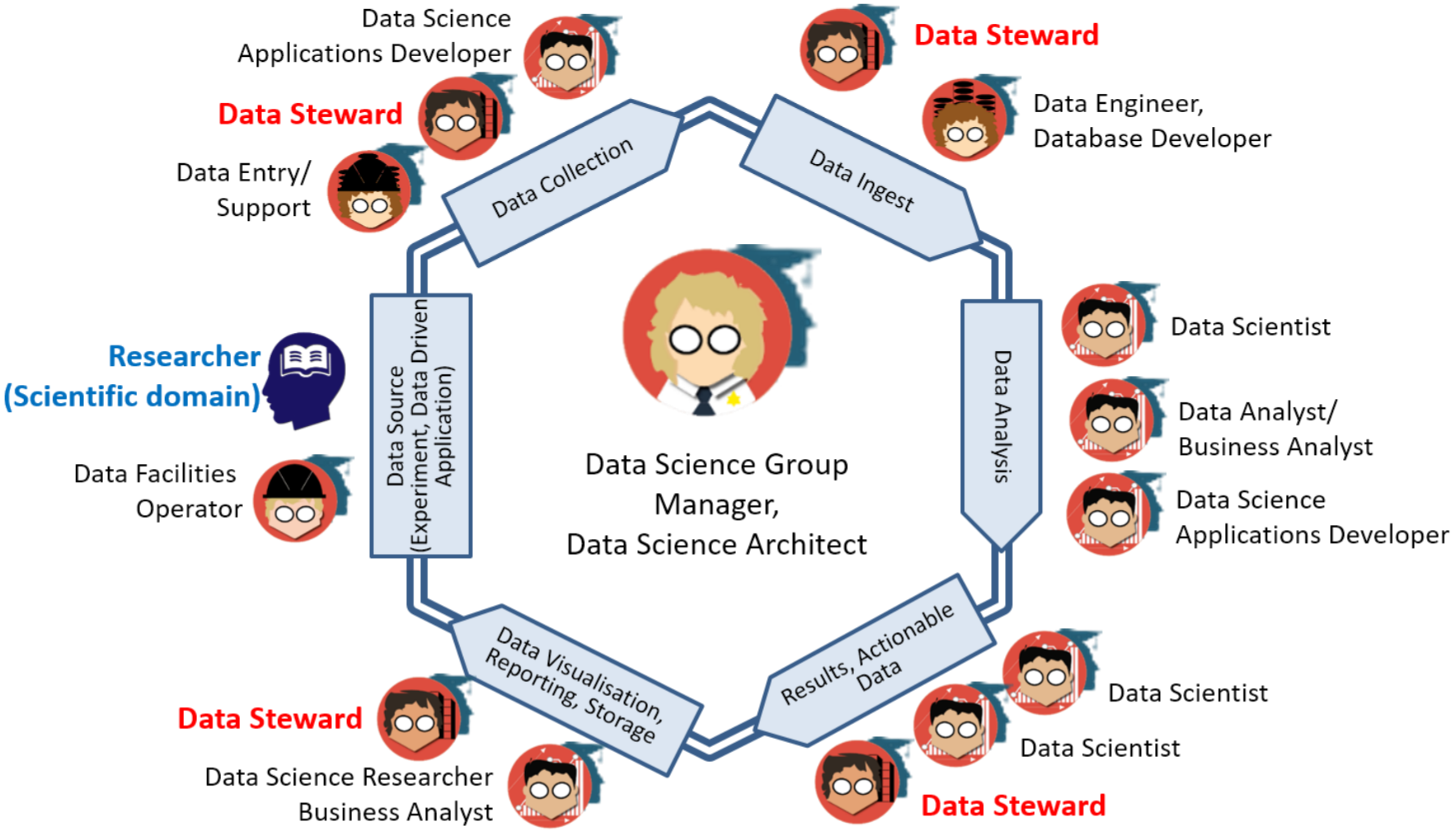
From Competence gap to Proficiency Level and Curriculum Timing – In progress



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



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, data librarians, 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.

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

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)

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

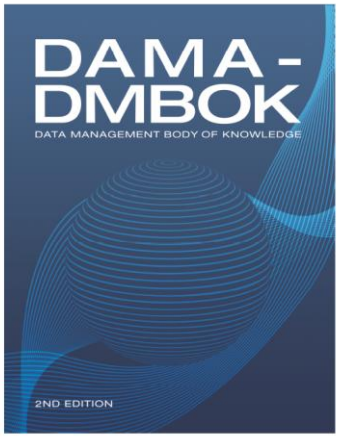
(13) Data Management Plan

(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



Technics Publications
BASKING RIDGE, NEW JERSEY

Data Governance and Stewardship

Definition: The exercise of authority, control, and shared decision-making (planning, monitoring, and enforcement) over the management of data assets.

Goals:
1. Enable an organization to manage its data as an asset.
2. Define, approve, communicate, and implement principles, policies, procedures, metrics, tools, and responsibilities for data management.
3. Monitor and guide policy compliance, data usage, and management activities.

Business Drivers

Inputs:
- Business Strategies & Goals
- IT Strategies & Goals
- Data Management and Data Strategies
- Organization Policies & Standards
- Business Culture Assessment
- Data Maturity Assessment
- IT Practices
- Regulatory Requirements

Activities:
1. **Define Data Governance for the Organization (P)**
   1. Develop Data Governance Strategy
   2. Perform Readiness Assessment
   3. Perform Discovery and Business Alignment
   4. Develop Organizational Touchpoints
2. **Define the Data Governance Strategy (P)**
   1. Define the Data Governance Operating Framework
   2. Develop Goals, Principles, and Policies
   3. Underwrite Data Management Projects
   4. Engage Change Management
   5. Engage in Issue Management
   6. Assess Regulatory Compliance Requirements
3. **Implement Data Governance (O)**
   1. Sponsor Data Standards and Procedures
   2. Develop a Business Glossary
   3. Co-ordinate with Architecture Groups
   4. Sponsor Data Asset Valuation
4. **Embed Data Governance (C,O)**

Deliverables:
- Data Governance Strategy
- Data Strategy
- Business / Data Governance Strategy Roadmap
- Data Principles, Data Governance Policies, Processes
- Operating Framework
- Roadmap and Implementation Strategy
- Operations Plan
- Business Glossary
- Data Governance Scorecard
- Data Governance Website
- Communications Plan
- Recognized Data Value
- Maturing Data Management Practices

Suppliers:
- Business Executives
- Data Stewards
- Data Owners
- Subject Matter Experts
- Maturity Assessors
- Regulators
- Enterprise Architects

Participants:
- Steering Committees
- CIO
- CDO / Chief Data Stewards
- Executive Data Stewards
- Coordinating Data Stewards
- Business Data Stewards
- Data Governance Bodies
- Compliance Team
- DM Executives
- Change Managers
- Enterprise Data Architects
- Project Management Office
- Governance Bodies
- Audit
- Data Professionals

Consumers:
- Data Governance Bodies
- Project Managers
- Compliance Team
- DM Communities of Interest
- DM Team
- Business Management
- Architecture Groups
- Partner Organizations

Technical Drivers

Techniques:
- Concise Messaging
- Contact List
- Logo

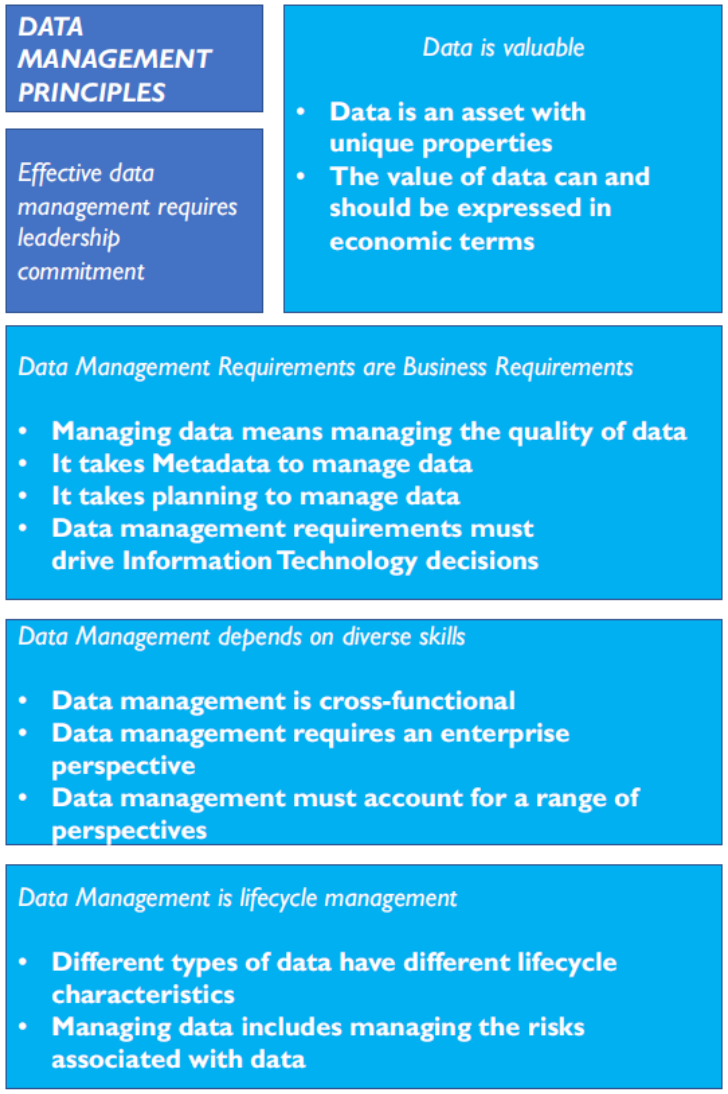
Tools:
- Websites
- Business Glossary Tools
- Workflow Tools
- Document Management Tools
- Data Governance Scorecards

Metrics:
- Compliance to regulatory and internal data policies.
- Value
- Effectiveness
- Sustainability

Scope of a Data Governance Programme

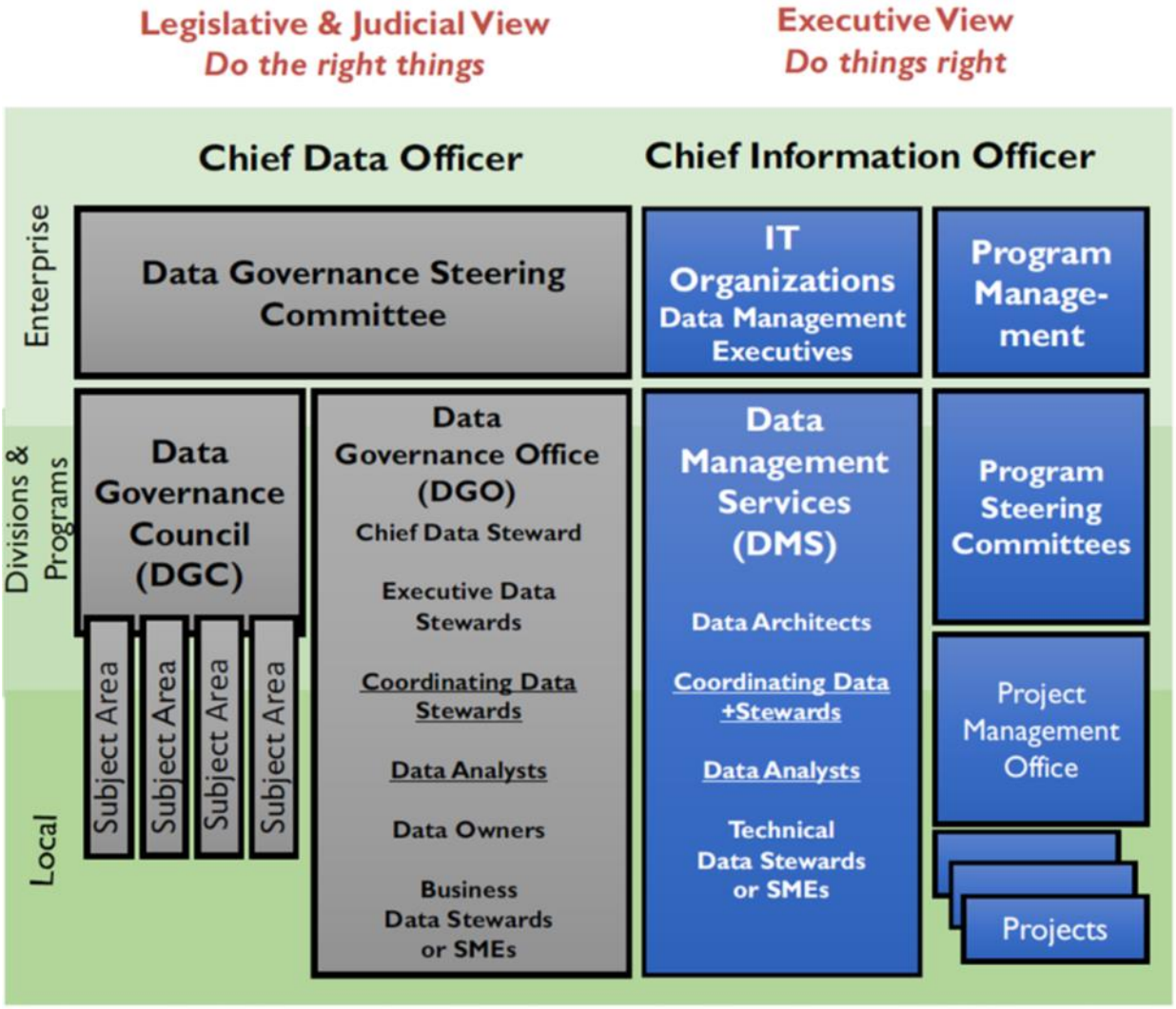
- Strategy
- Policy
- Standards and quality
- Oversight
- Compliance
- Issue management
- Data management projects
- Data asset valuation

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

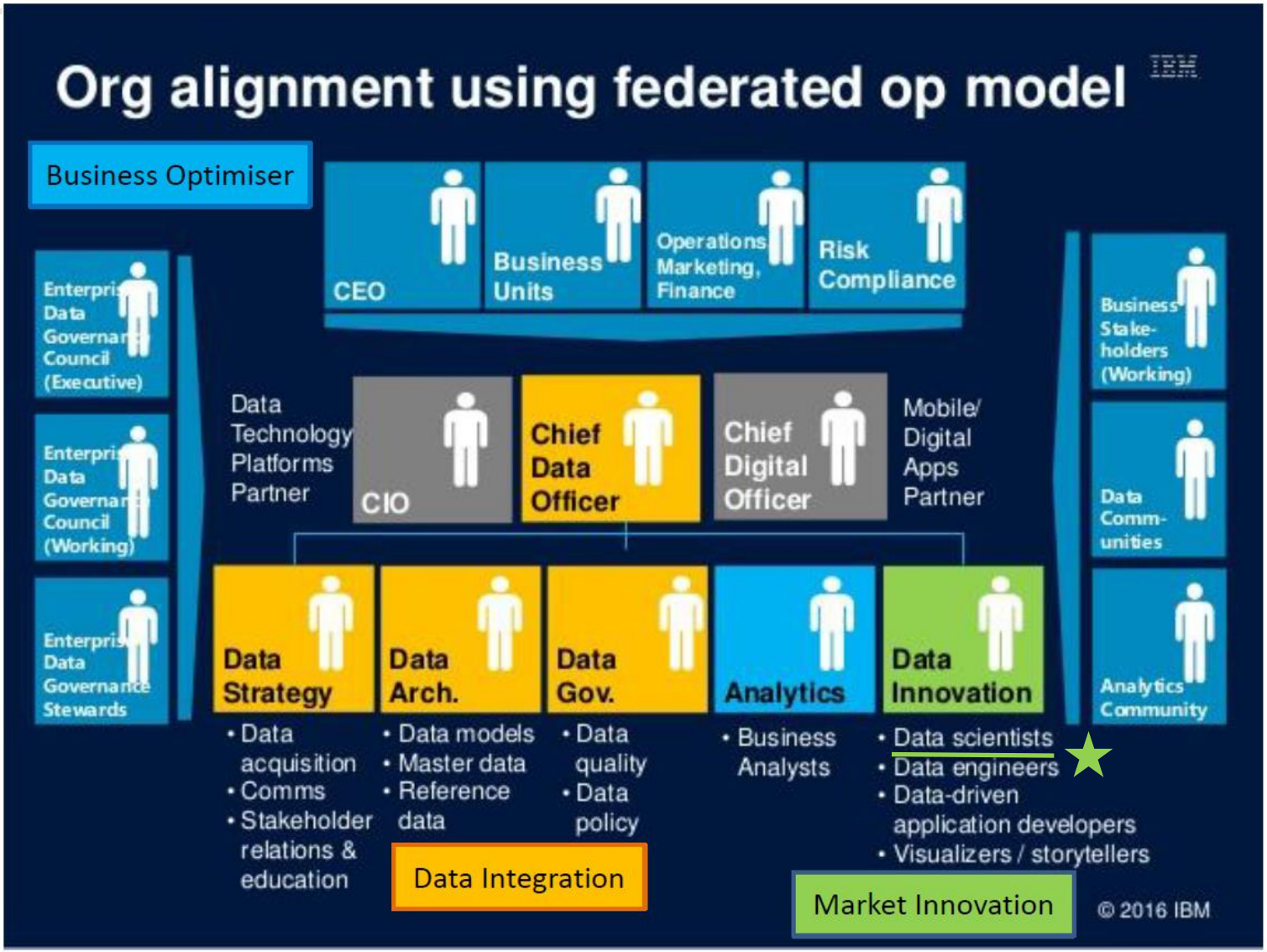
DMBOK: Data Governance Organisation Parts



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

EXAMPLE of IBM definition of new organisational roles

[ref] Cortnie Abercrombie, What CEOs want from CDOs and how to deliver on it (2016) [online] http://www.slideshare.net/IBMBDA/what-ceos-want-from-cdos-and-how-to-deliver-on-it



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)

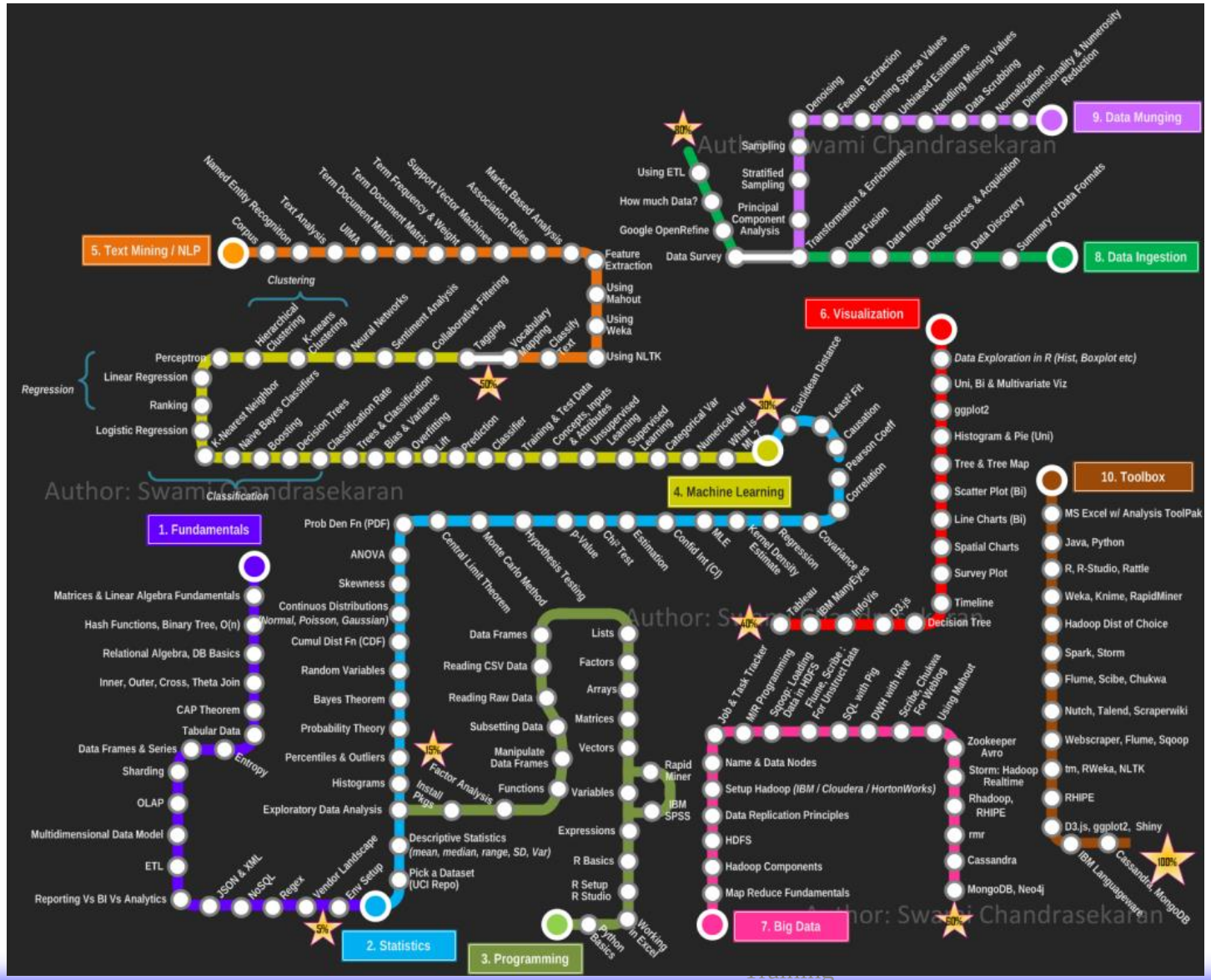
Discussion: How to become a Data Scientist

- A lot of information and different paths
- There are essential knowledge and competences
  - However most of them require strong background in mathematics, statistics, programming, infrastructure, etc.

Discussion: How to become a Data Scientist

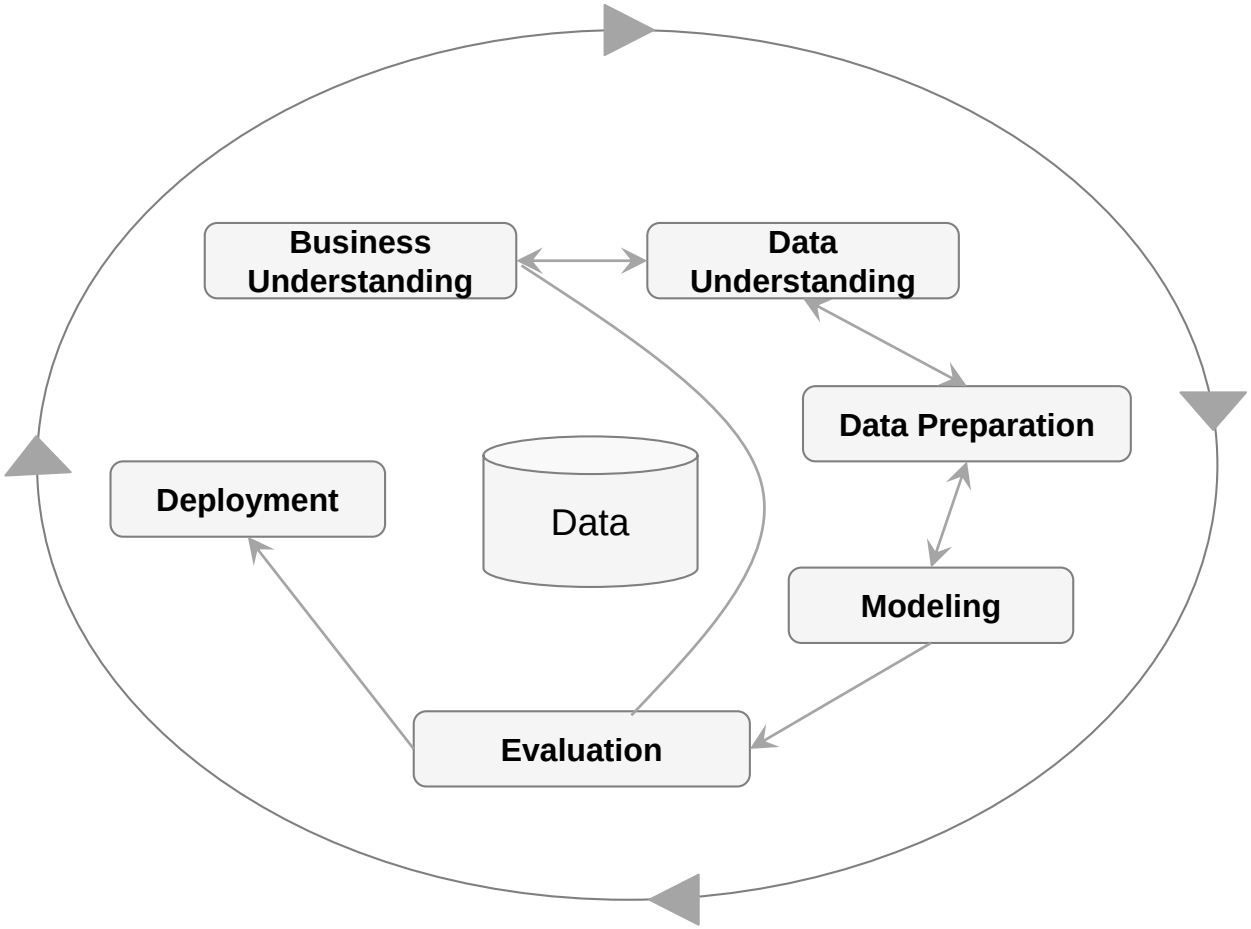
- Understand required Data Science and Analytics competences and skills
- Build your own learning path
  - Assess your knowledge and start from basics
  - Statistics is foundation of Data (Science) Analytics
    - Develop statistical/probabilistic thinking
    - Difference between Data Science and statistics
  - 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
  - Understand what company is actually looking for

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

CRISP DM process: Processes and Data Lifecycle

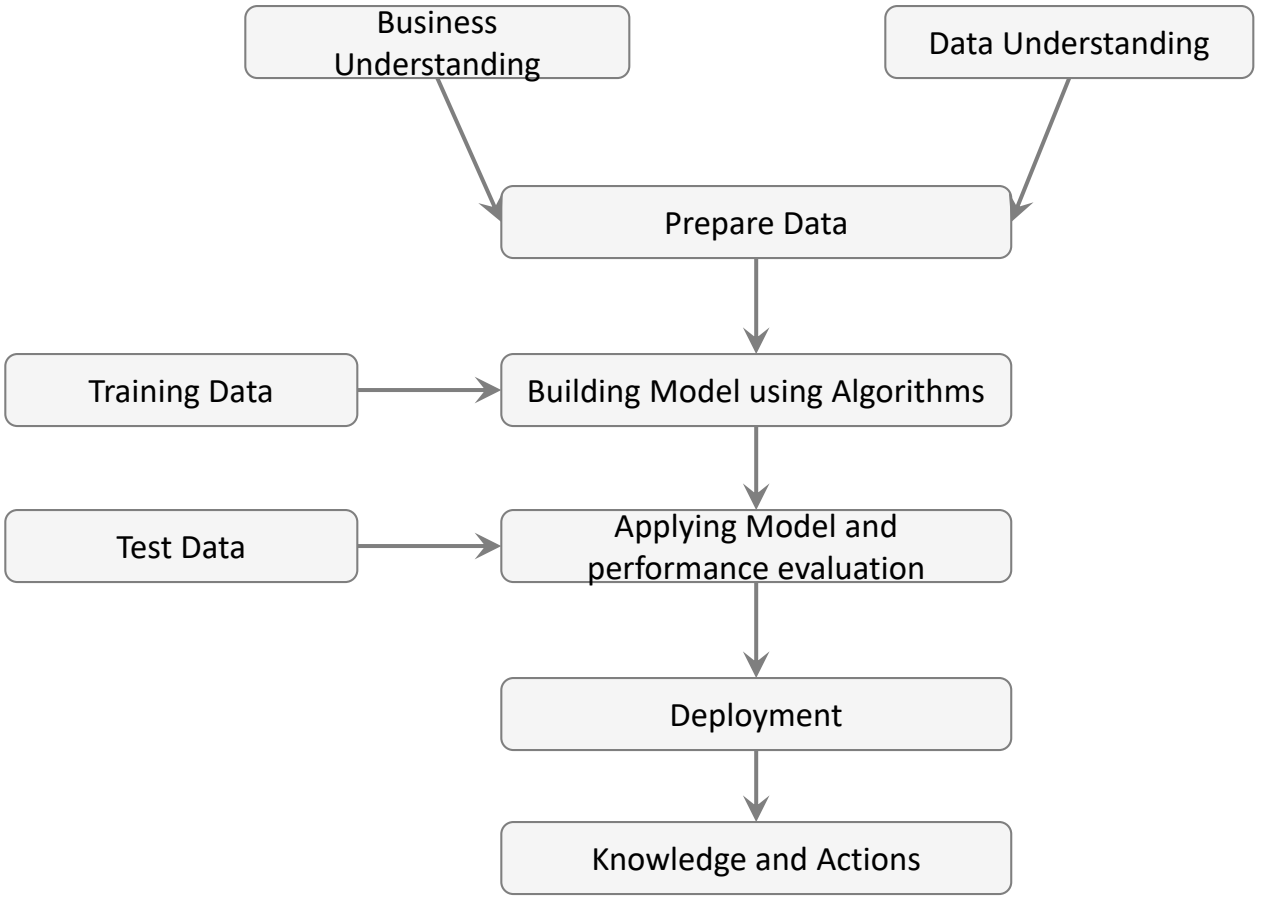


Cross Industry Standard Process for Data Mining (CRISP-DM) model and stages

- Business understanding
- Data Understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

All stages are iterative with the goal to achieve effectiveness for business decision making

Process of Data Analysis (based on CRISP-DM)



1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge

Online Educational and training resources

- LinkedIn Education – paid 279 USD/Yr
- Microsoft Learning (former Virtual Academy)
- AWS Learning
- IBM – in transition (after down since mid 2018)
- DataCamp paid 360 USD/Yr, 1 Mo trial, free starter python and R courses – growing popularity
- Coursera, Udacity
- Certification and training PMI, DAMA, IIBA, TDWI

Open Data and Educational Datasets

- Amazon Web Services (AWS)
- Google
- Microsoft Azure

- Github Open Data - https://github.com/collections/open-data
- Kaggle
- KDNuggets – community forum and blog
- Stackoverflow – community Q&A forum
- others

Links to EDISON Resources

- EDISON Data Science Framework Release 3 (EDSF)
  https://github.com/EDISONcommunity/EDSF

Component EDSF documents

CF-DS – Data Science Competence Framework
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_CF-DS-release3-v09.pdf

DS-BoK – Data Science Body of Knowledge
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DS-BoK-release3-v04.pdf

MC-DS – Data Science Model Curriculum
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_MC-DS-release3-v04.pdf

DSPP – Data Science Professional profiles
https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DSPP-release3-v05.pdf

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

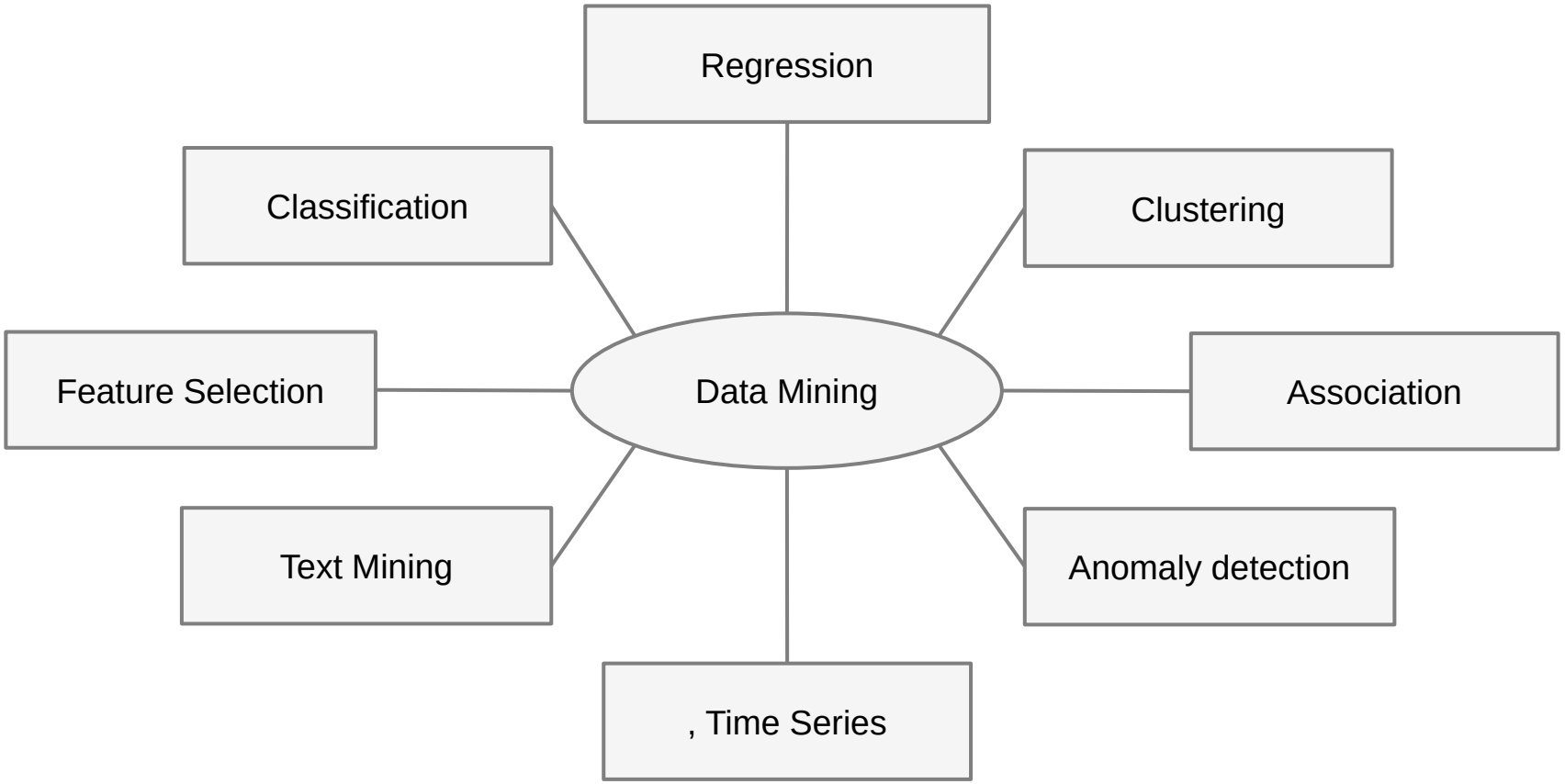
- EDISON project website - old domain ***edison-project.eu*** expired: Legacy information to be moved to http://edison-project.net/

Additional materials

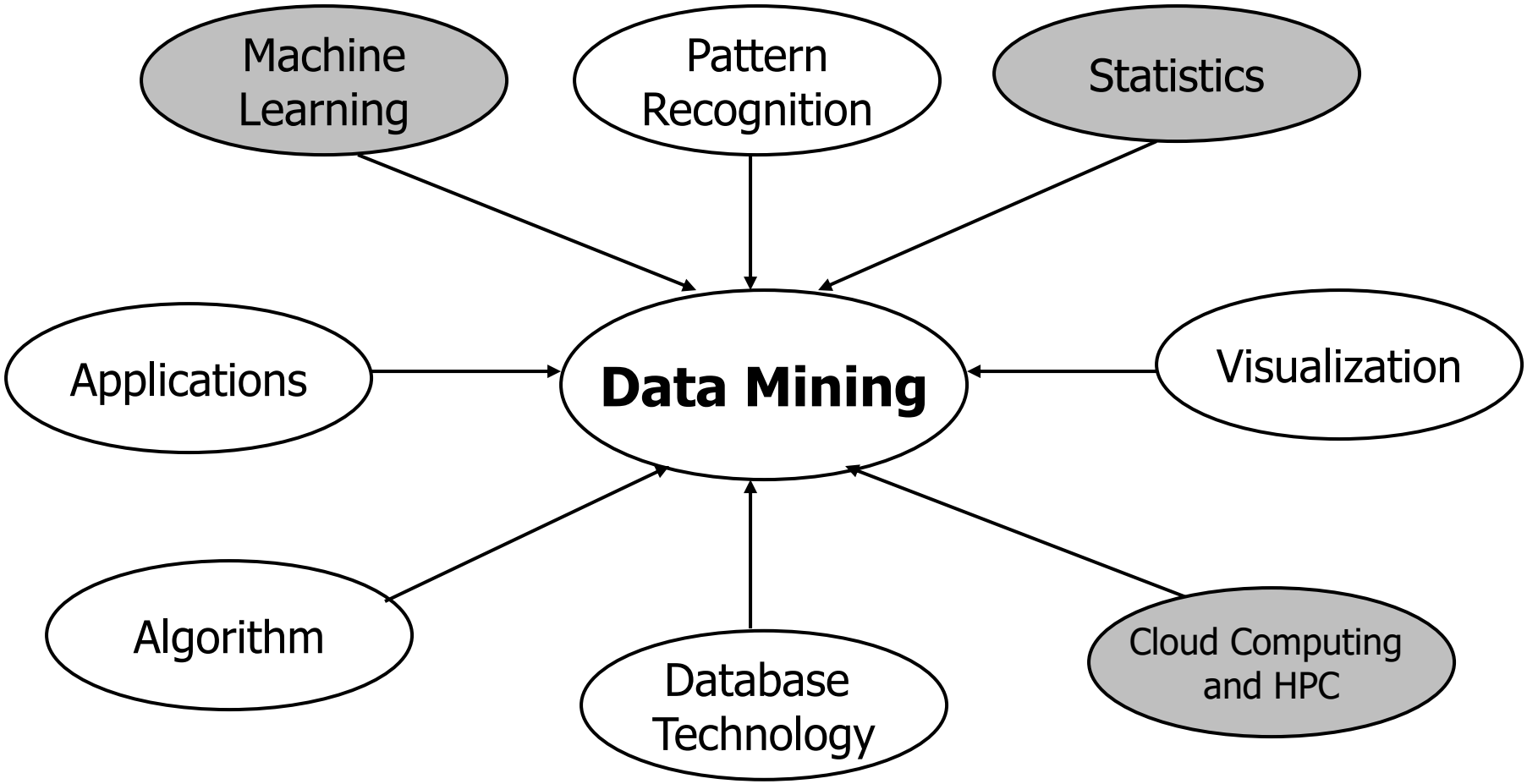
Data Science and Data Mining

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems

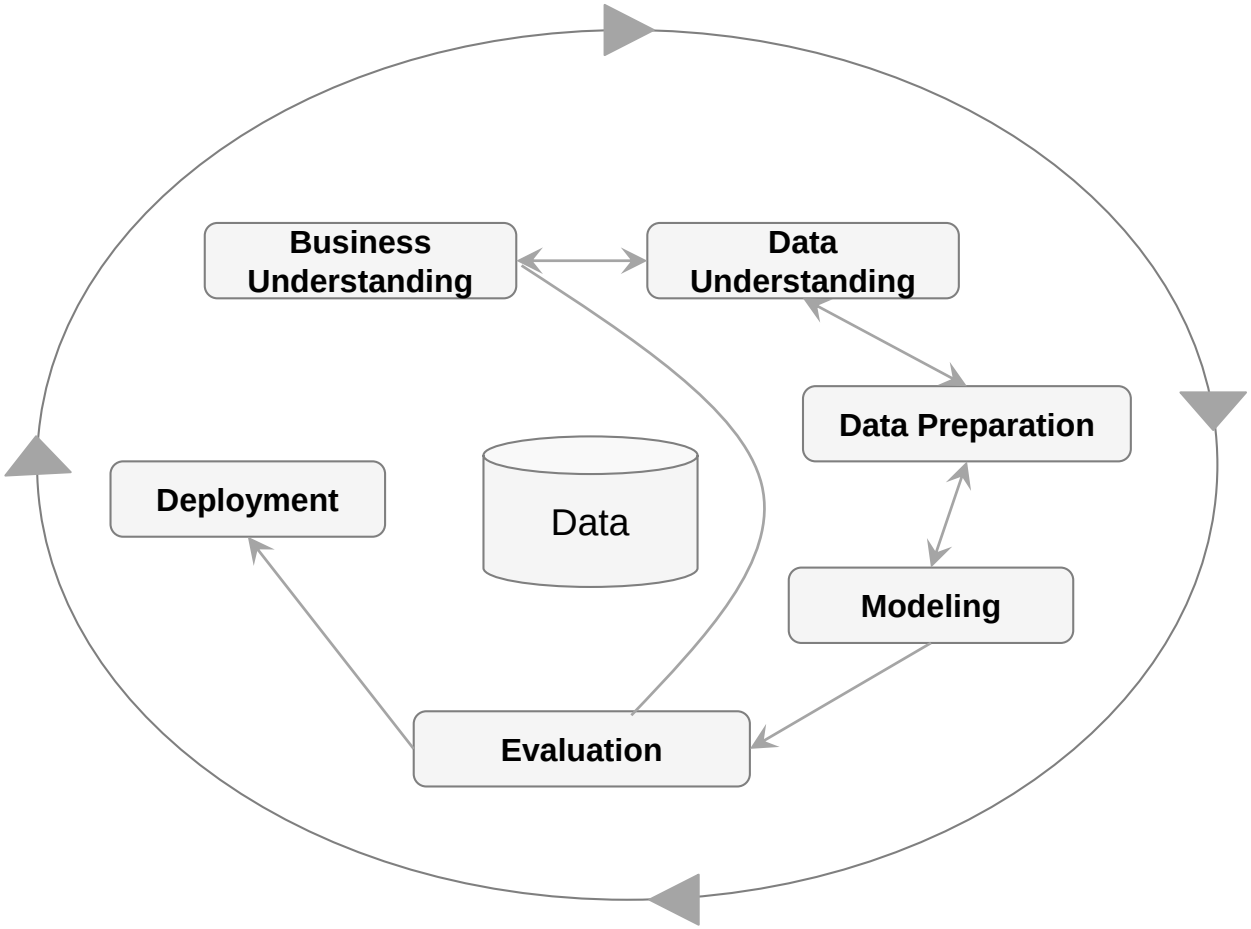
Types of Data Mining (branch of Data Analysis)



Data Mining: Confluence of Multiple Disciplines



CRISP DM process: Processes and Data Lifecycle

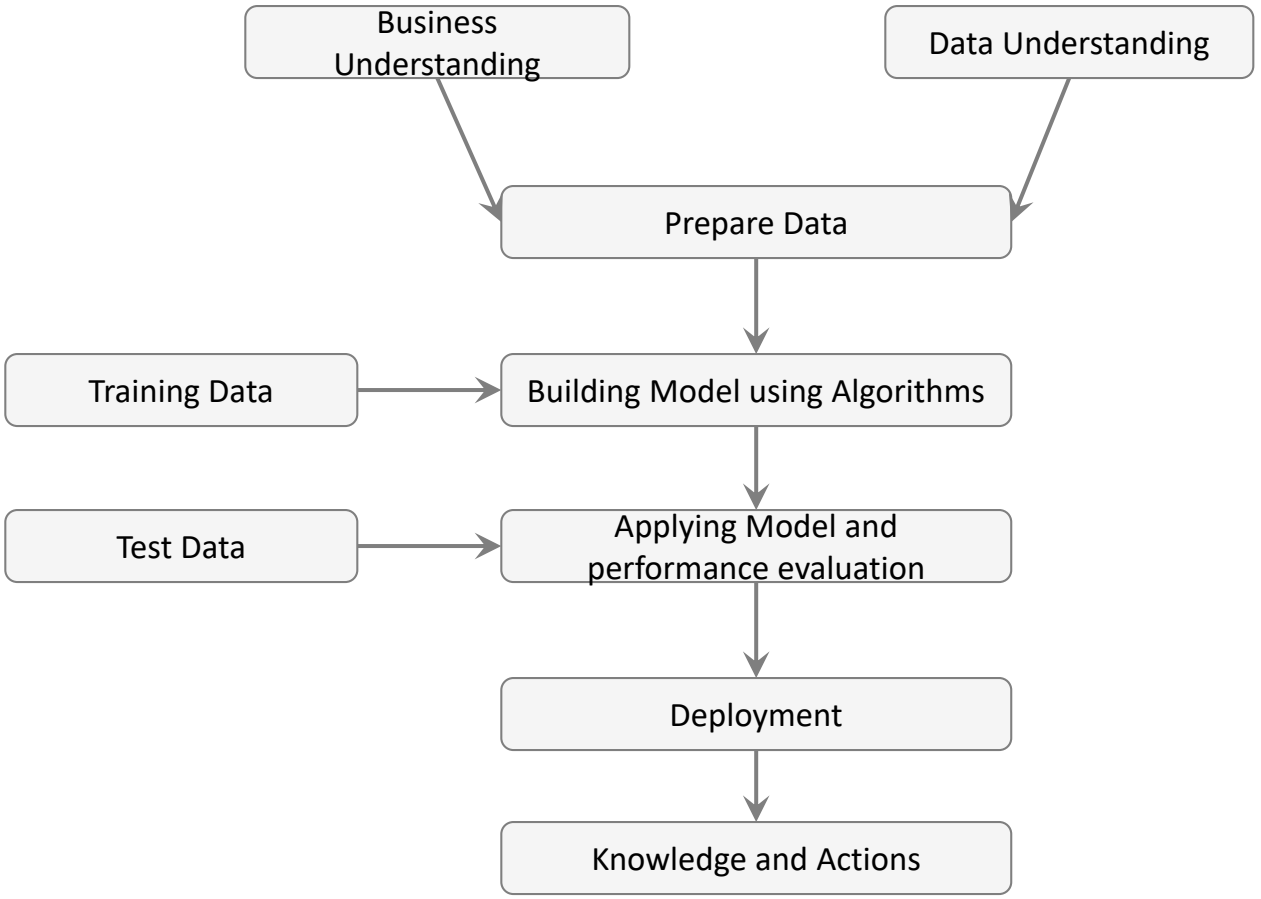


Cross Industry Standard Process for Data Mining (CRISP-DM) model and stages

- Business understanding
- Data Understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

All stages are iterative with the goal to achieve effectiveness for business decision making

Process of Data Analysis (based on CRISP-DM)



1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge