Becoming a Data Scientist and Educating Data Scientists:

Practical recommendations to develop Data Science and Analytics related competences and professional skills

EDISON
building the data science profession

Yuri Demchenko, University of Amsterdam

EDISON Project and Initiative

9 April 2018, Kiev

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

Grant 675419 (INFRASUPP-4-2015: CSA)

Outline

Part 1

- Background: Data driven research and demand for new skills
  - Foundation, recent reports, studies and facts

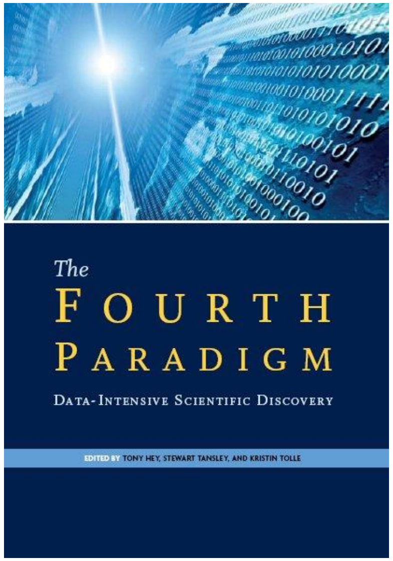
Part 2

- EDISON Data Science Framework (EDSF)
  - Data Science competences and skills
  - Essential Data Scientist professional skills: Thinking and doing like Data Scientist
- Data Science Professional Profiles
- Data Science Body of Knowledge and Model Curriculum

Pat 3

- Use of EDSF and Example curricula
  - Competences assessment
  - Building Data Science team
- Discussion

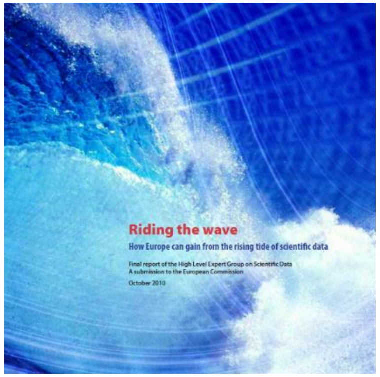
Visionaries and Drivers:
Seminal works, High level reports, Activities


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

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
A submission to the European Commission
October 2010

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

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)

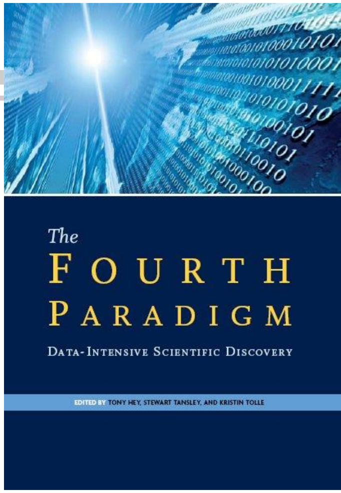
https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

Realising
the European
Open Science Cloud
First report and recommendations
of the Commission High Level Expert Group
on the European Open Science Cloud

Emergence of Cognitive Technologies

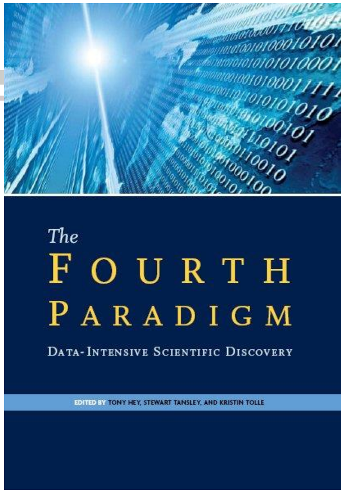
(IBM Watson, Cortana and others)

The Fourth Paradigm of Scientific Research


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

1. Theory, hypothesis and logical reasoning
2. Observation or Experiment, e.g.
   - Newton observed apples falling to design his theory of mechanics
   - Gallileo Galilei made experiments with falling objects from the Pisa leaning tower
3. Simulation of theory or model
   - Digital simulation can prove theory or model
4. Data-driven Scientific Discovery (aka Data Science)
   - More data beat hypothesized theory
   - e-Science as computing and Information Technologies empowered science
5. Computer-human - driven science?
   - Machine discovers new patterns and formulates hypothesis in one or multiples knowledge spaces
   - Scientist validates and designs additional texts or experiments

The Fourth Paradigm of Scientific Research


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

0. Belief and religion – which are actually observation based
1. Theory, hypothesis and logical reasoning
2. Observation or Experiment, e.g.
   - Newton observed apples falling to design his theory of mechanics
   - Gallileo Galilei made experiments with falling objects from the Pisa leaning tower
3. Simulation of theory or model
   - Digital simulation can prove theory or model
4. Data-driven Scientific Discovery (aka Data Science)
   - More data beat hypothesized theory
   - e-Science as computing and Information Technologies empowered science
5. Computer-human - driven science?
   - Machine discovers new patterns and formulates hypothesis in one or multiples knowledge spaces
   - Scientist validates and designs additional texts or experiments

EU Specific Drivers and Recommendations

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]

Realising the European Open Science Cloud

First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud

- **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/

International and EU studies on data-driven skills

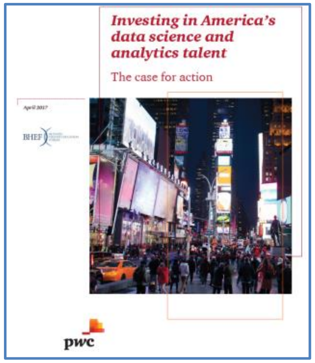
Industry reports on Data Science Analytics and Data enabled skills demand

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

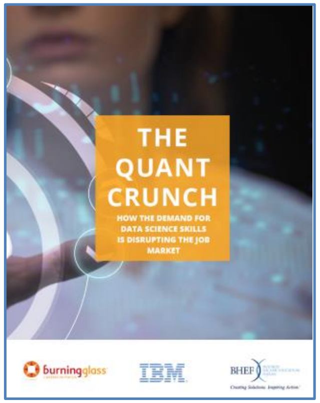
- PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017)
  - http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent
  - 2.35 mln postings, 23% Data Scientist, 67% DSA enabled jobs
  - DSA enabled jobs growing at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

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

IDC
Analyze the Future
European Data Market
SMART 2013/0063
Final Report
IDC
Analyze the Future
OPEN EVIDENCE
February 1st 2017

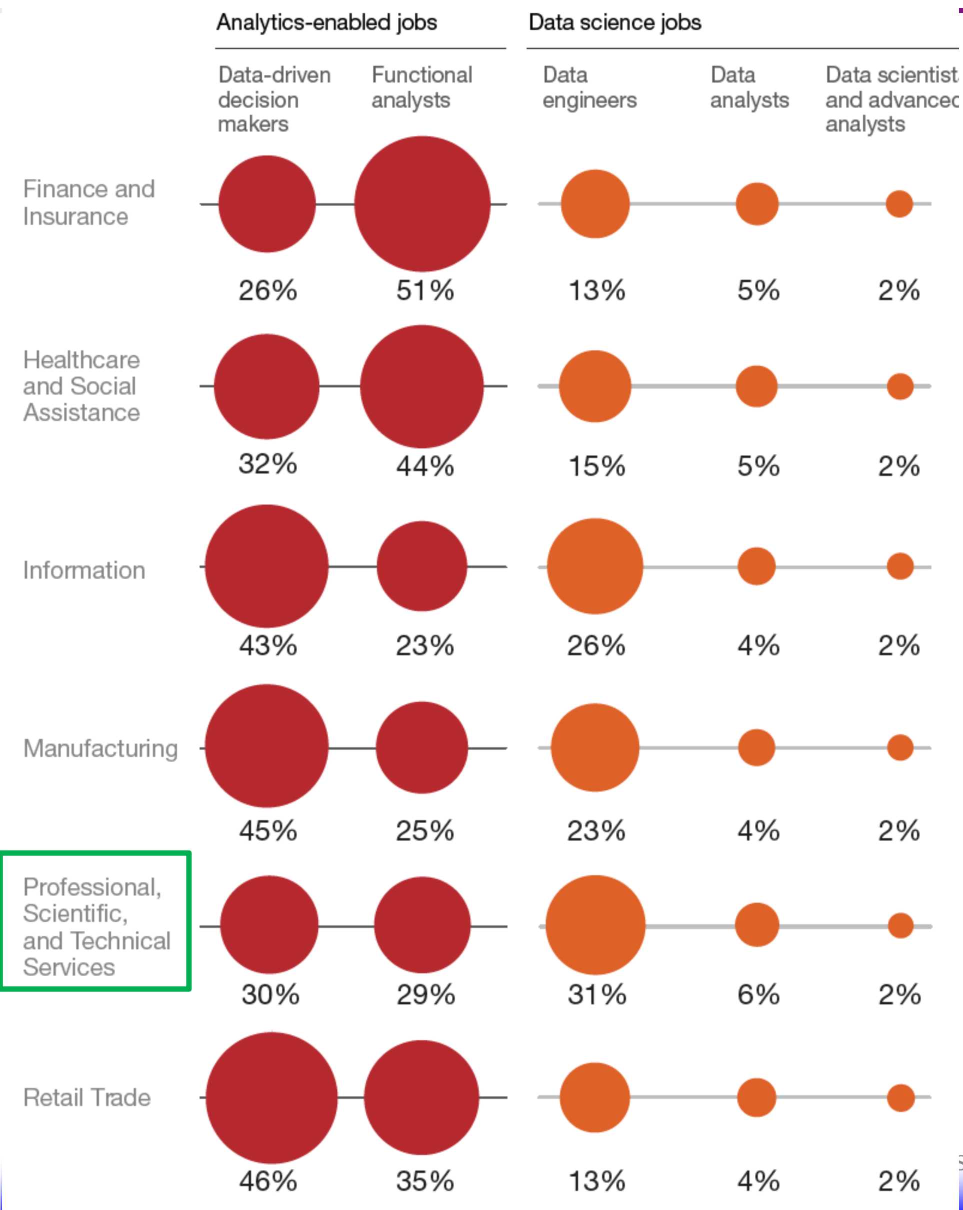

Investing in America’s data science and analytics talent
The case for action
April 2017
BHEF
pwc

Citing EDISON and EDSF


THE QUANT CRUNCH
HOW THE DEMAND FOR DATA SCIENCE SKILLS IS DISRUPTING THE JOB MARKET
burningglass
IBM
BHEF

Influenced by EDISON

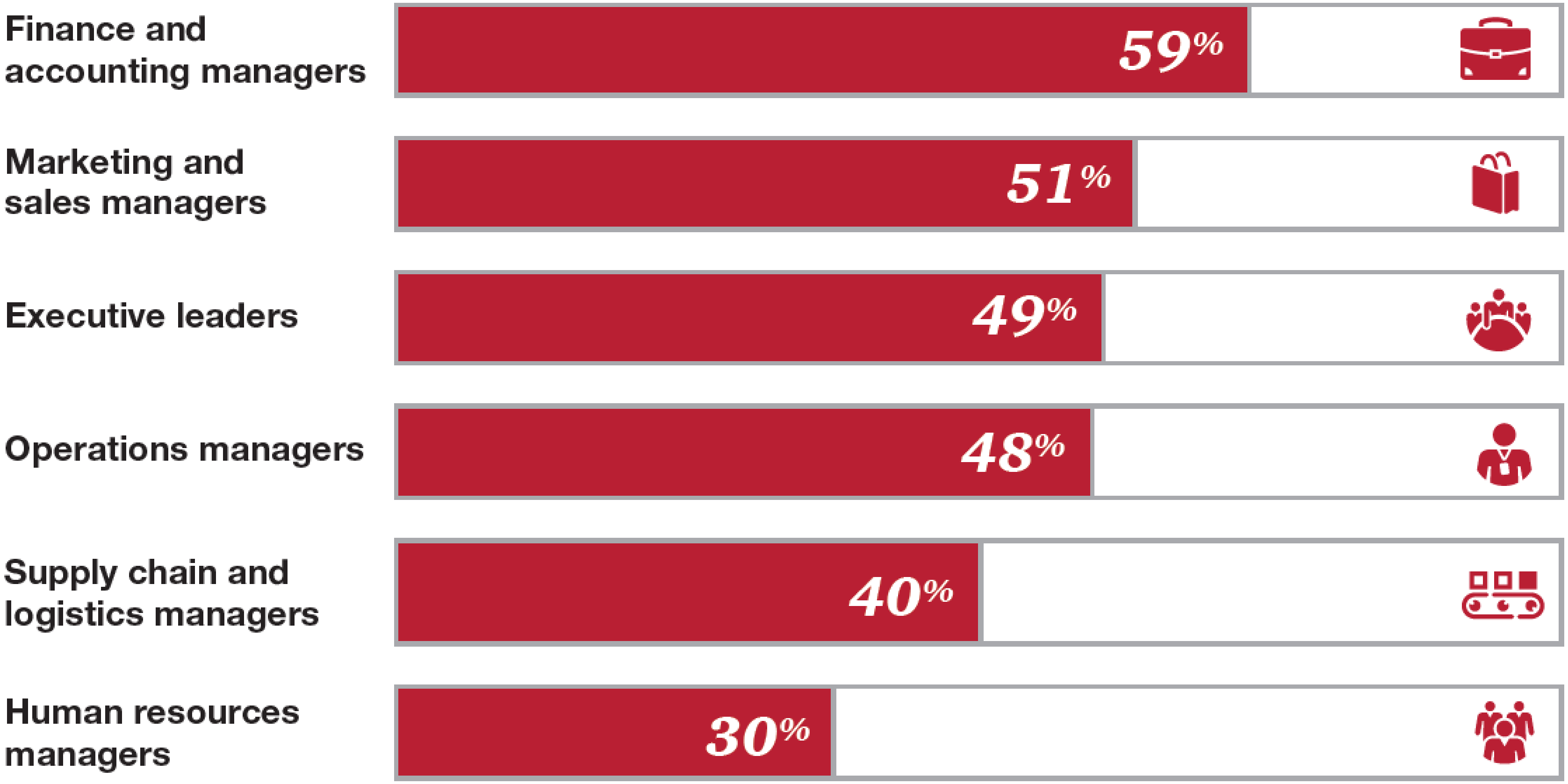
PwC&BHEF: Demand for DSA enabled jobs



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

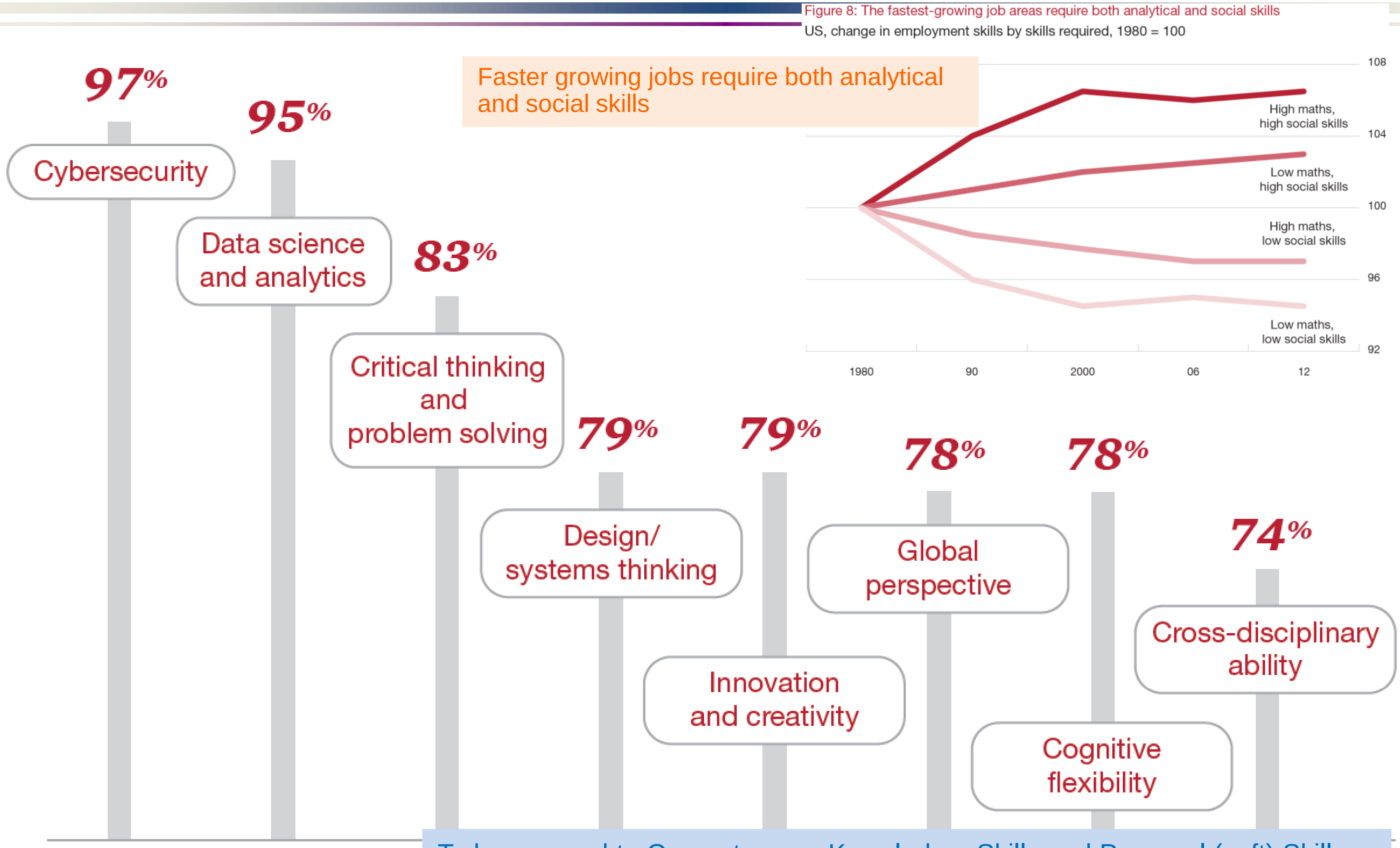


Figure 8: The fastest-growing job areas require both analytical and social skills

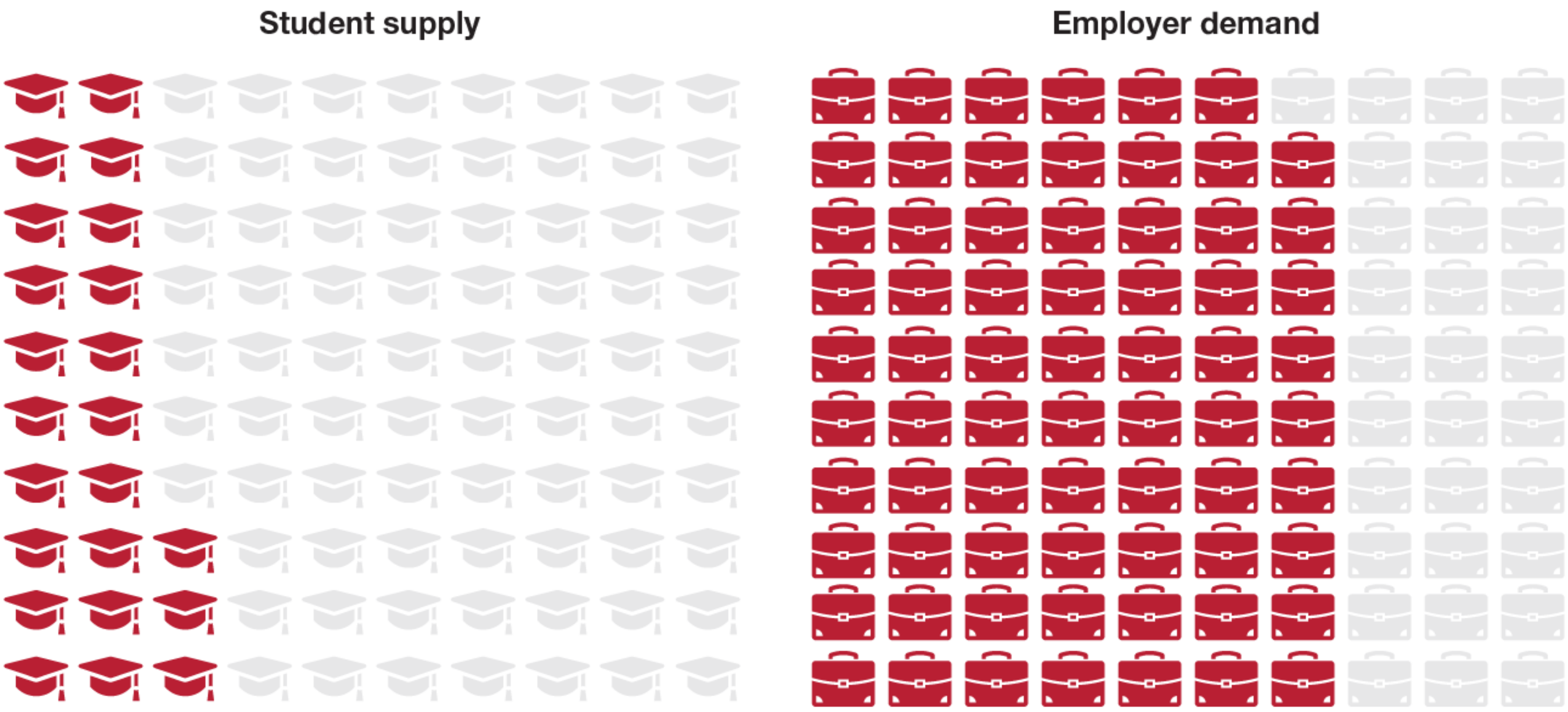
US, change in employment skills by skills required, 1980 = 100

Faster growing jobs require both analytical and social skills

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

Source: Business Roundtable (2017).

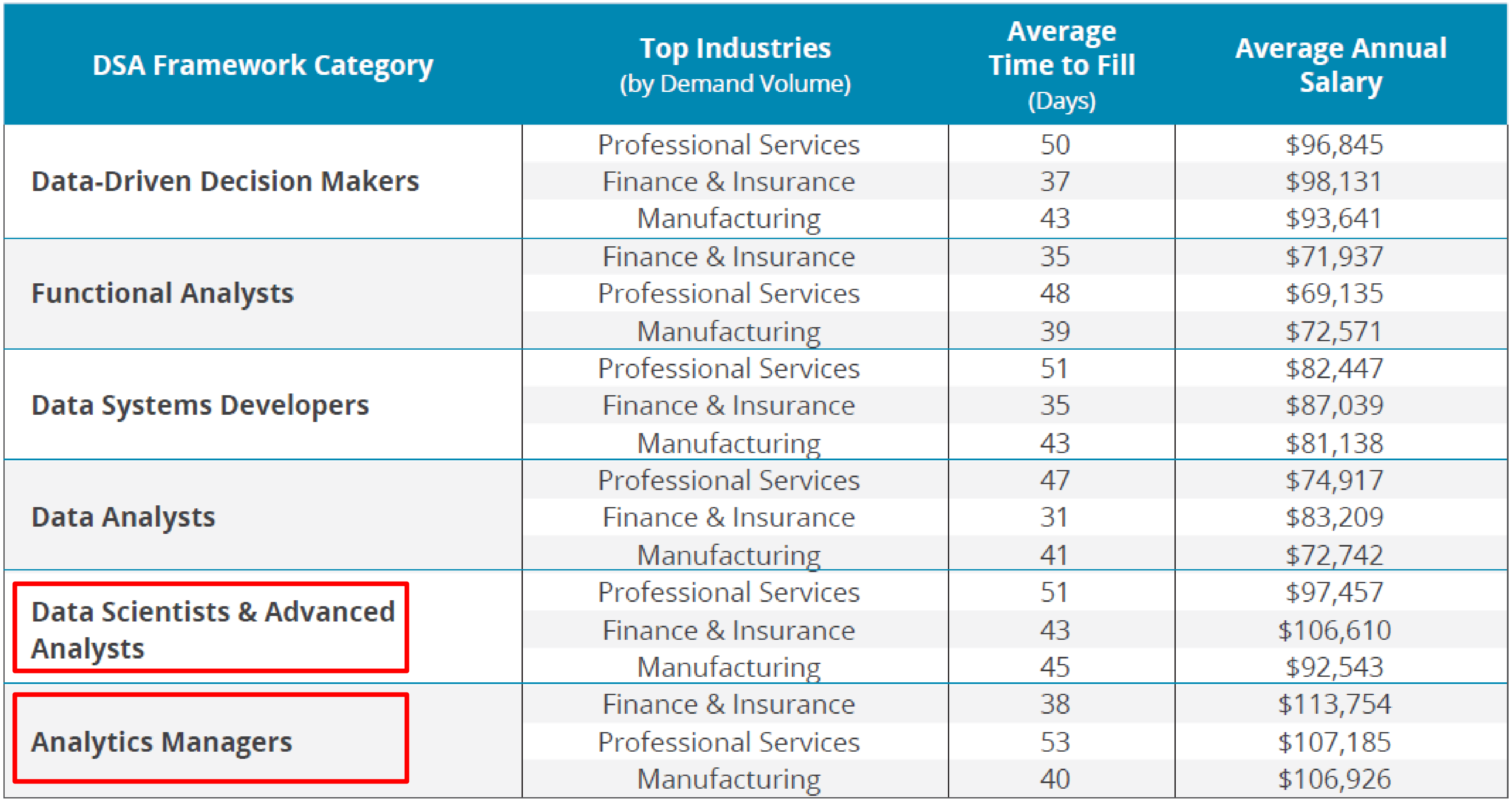
PwC&BHEF: Data Science and Analytics skills, by 2021: The supply-demand challenge



23% of educators say all graduates will have data science and analytics skills

69% of employers say they will prefer job candidates with these skills over ones without

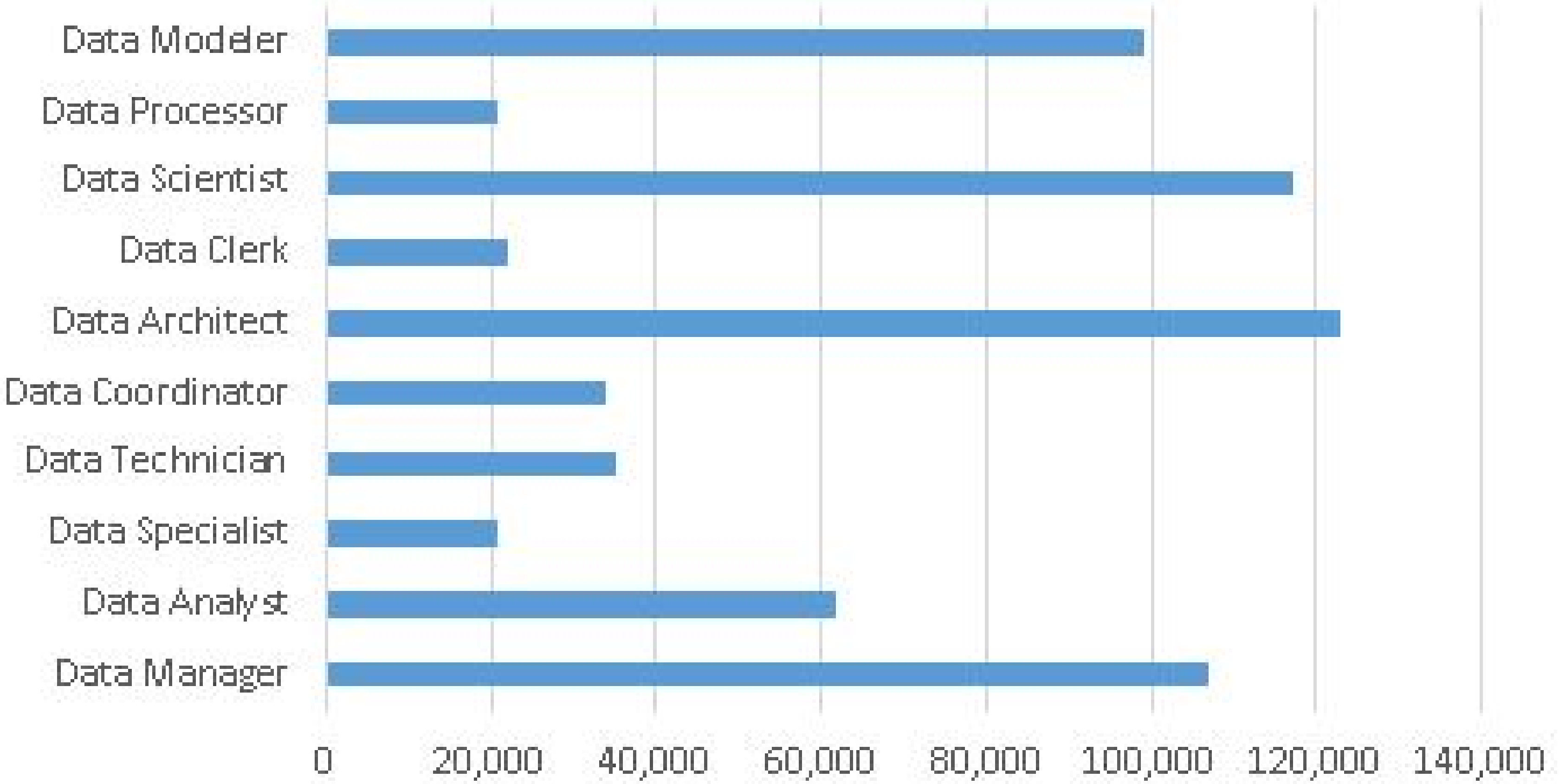
IBM&BGT: DSA Jobs Time to Fill and Salary (2016-2017)



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

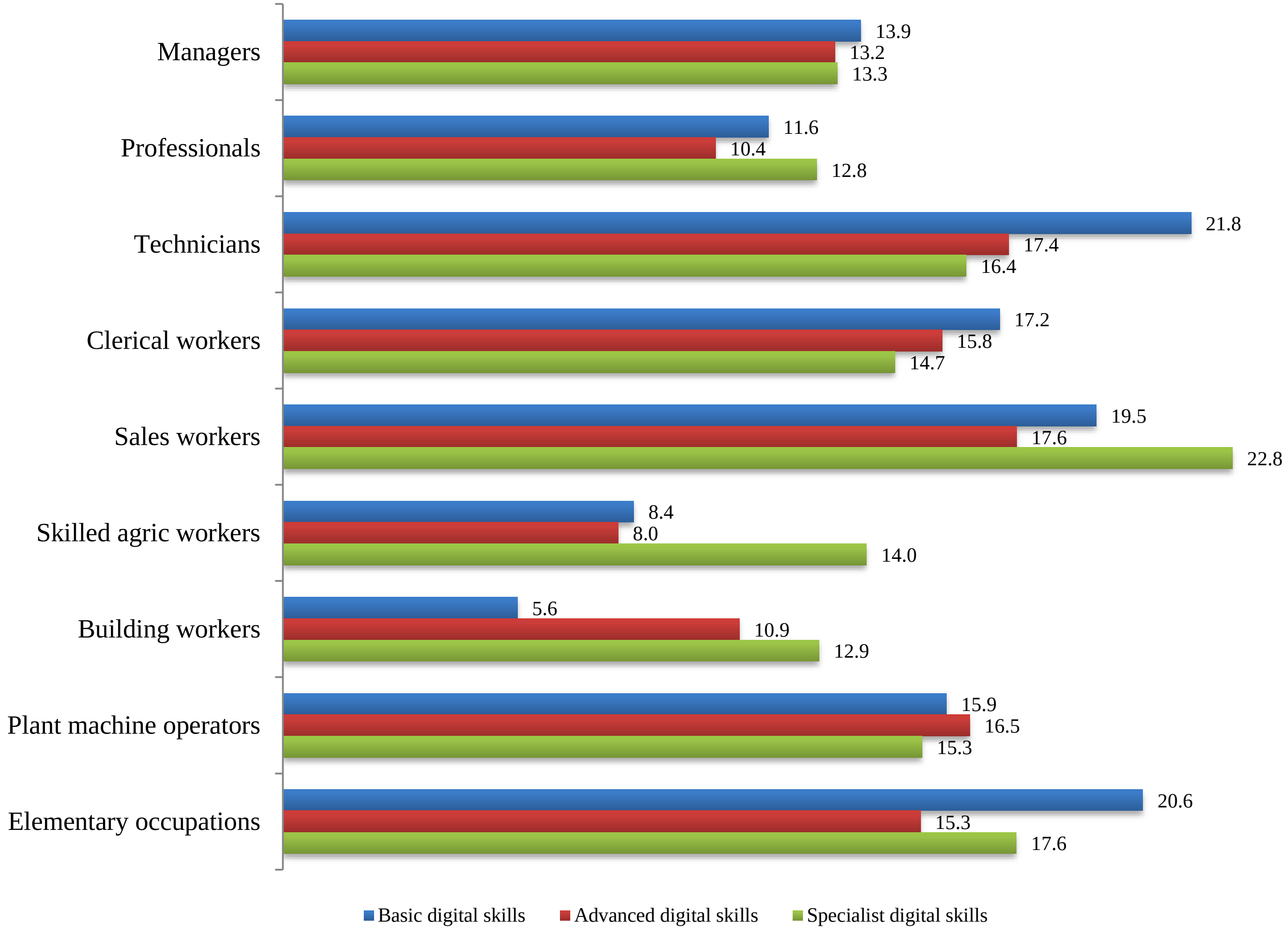
- On average, DSA jobs in Professional Services remain open for 53 days, eight days longer than the overall DSA average. (IBM, BGT 2017 Study)

Closer look at Data related Jobs and Salaries (2016)



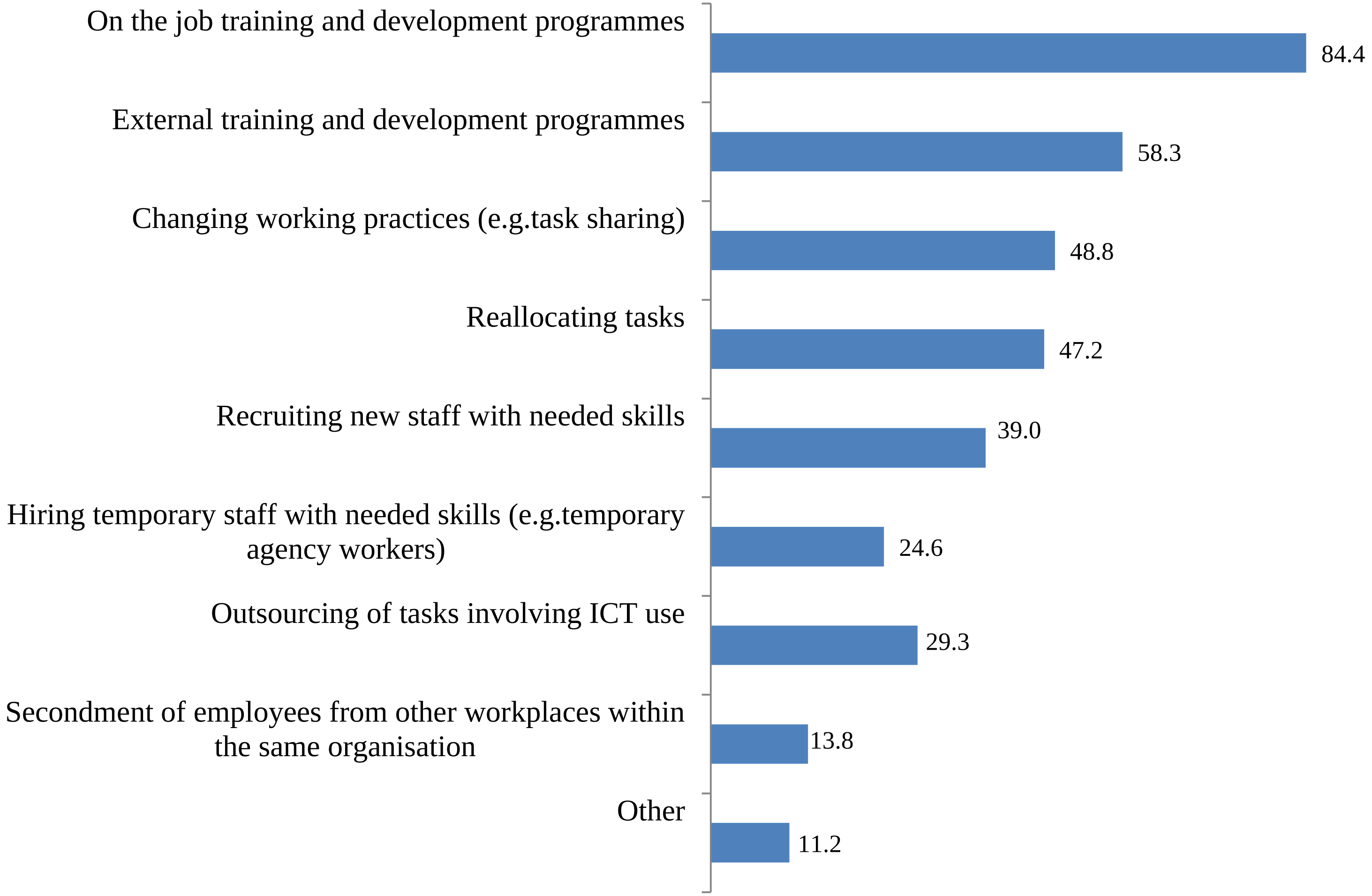
Source: The Job Market for Data Professionals, by Robert R Downs, SciDataCon2016
http://www.scidatacon.org/2016/sessions/98/poster/51/

Digital skills gaps density by occupation and type of digital skills, EU28 (%)



ICT for work: Digital skills in the workplace, Digital Single Market, Reports and studies, May 2017
https://ec.europa.eu/digital-single-market/en/news/ict-work-digital-skills-workplace

Workplaces reporting having taken action to tackle digital skill gaps by type of action undertaken, EU28 (% of workplaces with digital skill gaps which undertook actions)



ICT for work: Digital skills in the workplace, Digital Single Market, Reports and studies, May 2017
https://ec.europa.eu/digital-single-market/en/news/ict-work-digital-skills-workplace

OECD and UN on Digital Economy and Data Literacy

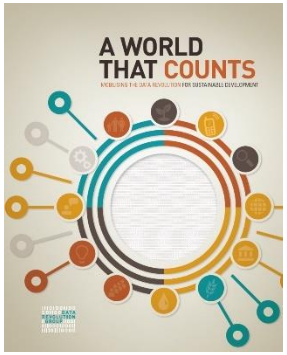
OECD (Organisation for Economic Coopration and Development)

- 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

www.pwc.com

Millennials at work

Reshaping the workplace

ial now o t, will world you ready?

pwc

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

- 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

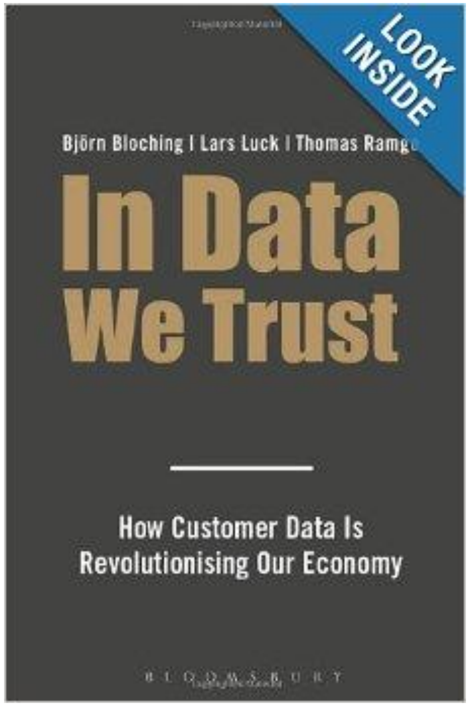
Data Driven Victories and Failures - Politics

Very high impact events and facts

- **US Election 2012** – Obama’s campaign and rise of Big Data analytics
  - Micro-targeting and Social Networks analysis
- **Brexit 2016**
  - “Data driven Brexit” – first serious ring for right use of Data Science technologies
- **US Election 2016**
  - Clinton’s campaign – “Data driven” but using only upper layer of Social Network (SN) web
  - Trump’s campaign – Targeting bottom SN web and “forgotten people not to be forgotten”
    - Matt Oczkowski, leader on Trump’s campaign: “If he was going to win this election, it was going to be because of a Brexit style mentality and a different demographic trend than other people were seeing.”
- France election 2017
  - Awakening

Data-Driven Brexit: A Wakeup Call for Analysts

By Barry Devlin, June 28, 2016



LOOK INSIDE
Björn Bloching | Lars Luck | Thomas Ramge
In Data We Trust
How Customer Data Is Revolutionising Our Economy
BLOOMSBURY

Book: In Data We Trust: How Customer Data is Revolutionising Our Economy (Aug 2012)

- A strategy for tomorrow's data world

Data-Driven Brexit: A Wakeup Call for Analysts

By Barry Devlin, June 28, 2016

Data-Driven Brexit: A Wakeup Call for Analysts

By Barry Devlin

June 28, 2016

The morning of Friday, June 24 saw red on the markets. Mainstream politicians were horrified. The great British public had delivered their verdict and it came as a shock. Why? Wasn't the polling data clear enough in the weeks before the vote? Hadn't the British people -- and, indeed, the world at large -- been shown all the facts needed to make rational, thoughtful decisions?

There are significant lessons for believers in data-driven business to learn from used for decision making before, during, and after the Brexit vote.

Consider first the shock and horror expressed on Friday morning. Polls over the

Data-Driven Brexit: A Wakeup Call for Analysts

SHARE WITH:

- Article "In Data we trust" by T.Edsall in The New York Times
- Multimillion-dollar contract for data management and collection services awarded May 1, 2013 to Liberty Work (for Republicans) to build advanced list of voters

- There are significant lessons for believers in data-driven business to learn from how data was and wasn't used for decision making before, during, and after the Brexit vote.
- Human attitude -- including emotion, intuition, and social empathy -- and motivation are at the heart of decision making and the action that follows
- Information will only be accepted when it conforms to preconceived notions. Expertise is not sufficient and, *in extremis*, will be dismissed with ridicule.

US elections 2016 and Data Analytics

- On-going scandal with Cambridge Analytica

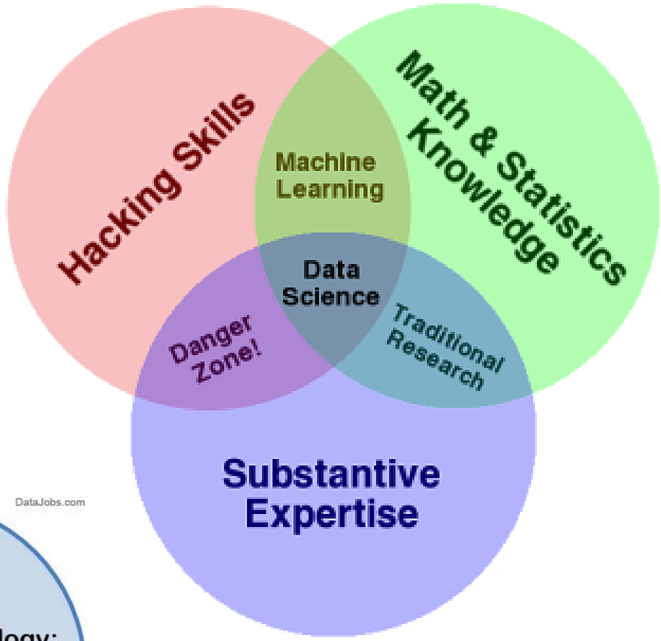
- Growing importance of ethical factor
  - Education is essential to tame new element/dimension of our life - Data

- Increasing impact of EU GDPR (General Data Protection Regulation) to be in force from 25 May 2016

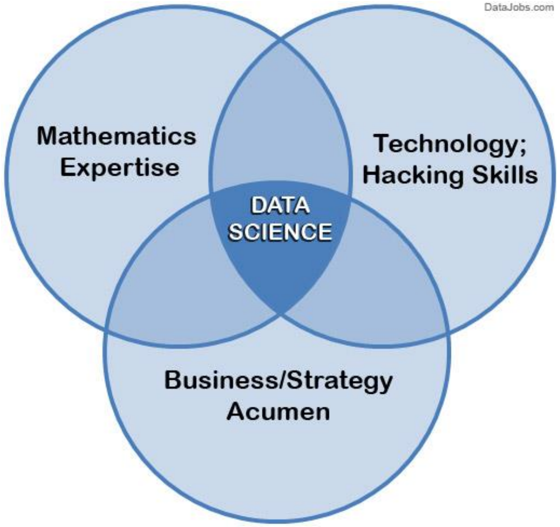
Challenge for Education:
Sustainable ICT and Data Skills Development

- 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

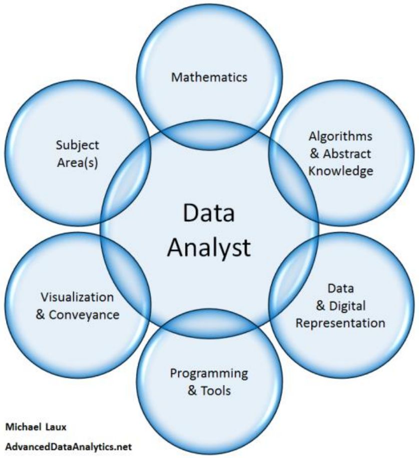
Data Scientist definitions: From Math to Hacking



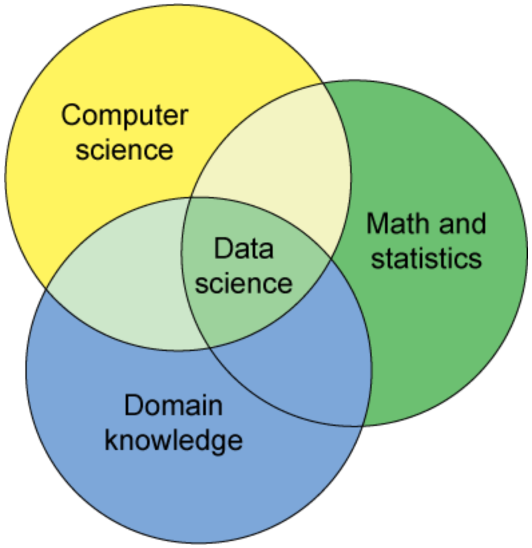
- Strongly depend on the background of the Data Scientist



DataJobs.com

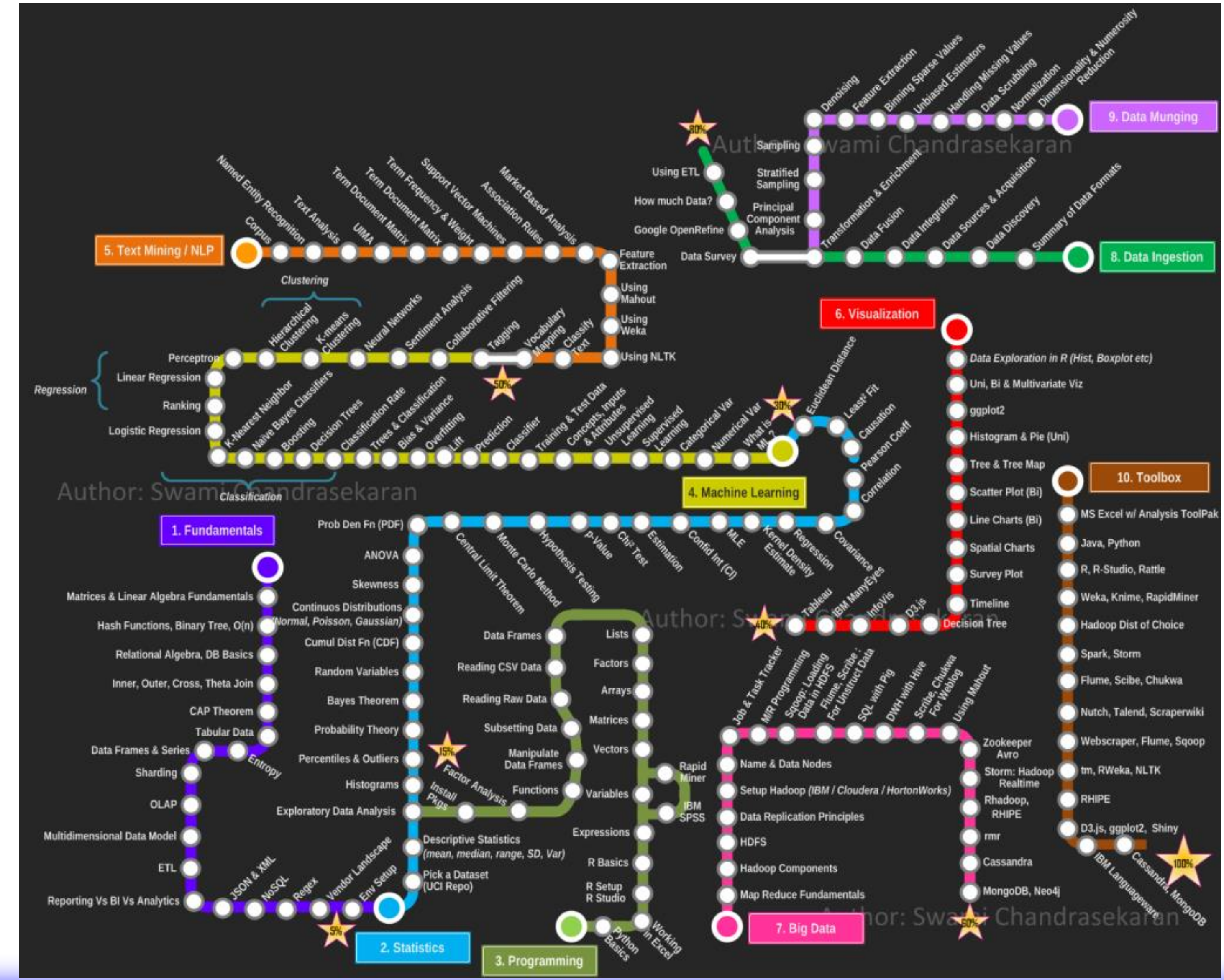

Mathematics
Subject Area(s)
Algorithms & Abstract Knowledge
Data Analyst
Visualization & Conveyance
Data & Digital Representation
Programming & Tools

Michael Laux
AdvancedDataAnalytics.net



Becoming a Data Scientist by Swami Chandrasekaran (2013)

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

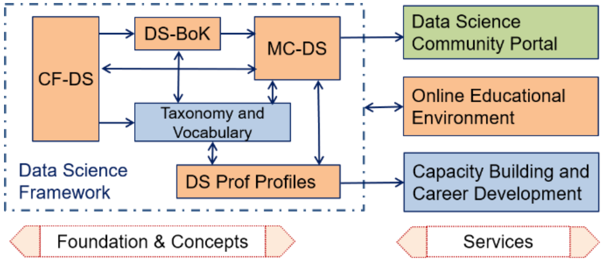


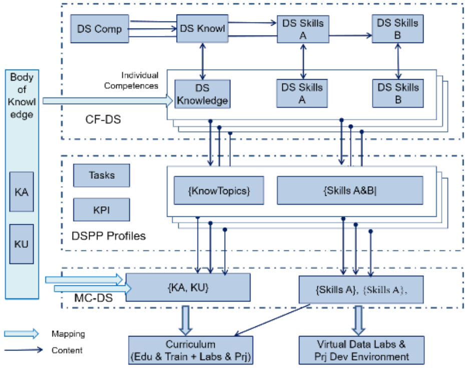
- Good and practical advice how to learn Data Science, step by step

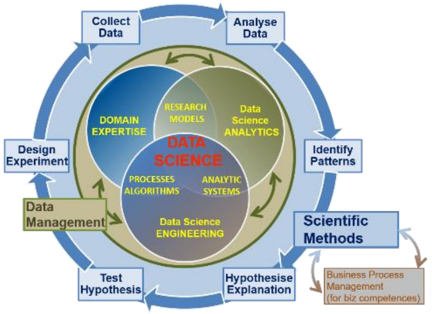
- Follow the route

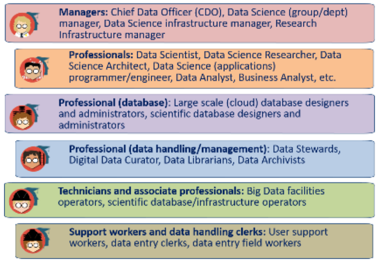
EDISON Products for Data Science Skills Management and Curriculum Design

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









Managers: Chief Data Officer (CDO), Data Science (group/dept) manager, Data Science infrastructure manager, Research Infrastructure manager

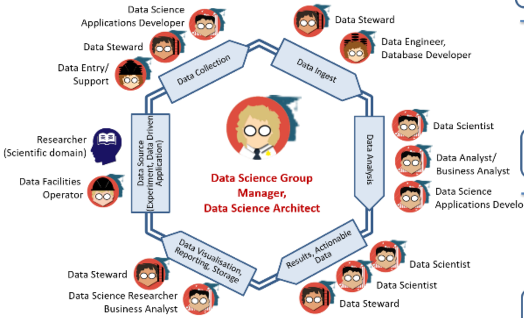
Professionals: Data Scientist, Data Science Researcher, Data Science Architect, Data Science (applications) programmer/engineer, Data Analyst, Business Analyst, etc.

Professional (database): Large scale (cloud) database designers and administrators, scientific database designers and administrators

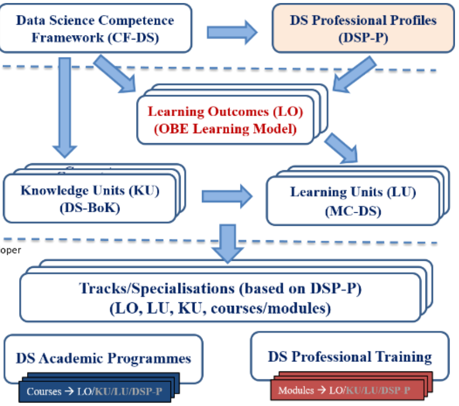
Professional (data handling/management): Data Stewards, Digital Data Curator, Data Librarians, Data Archivists

Technicians and associate professionals: Big Data facilities operators, scientific database/infrastructure operators

Support workers and data handling clerks: User support workers, data entry clerks, data entry field workers

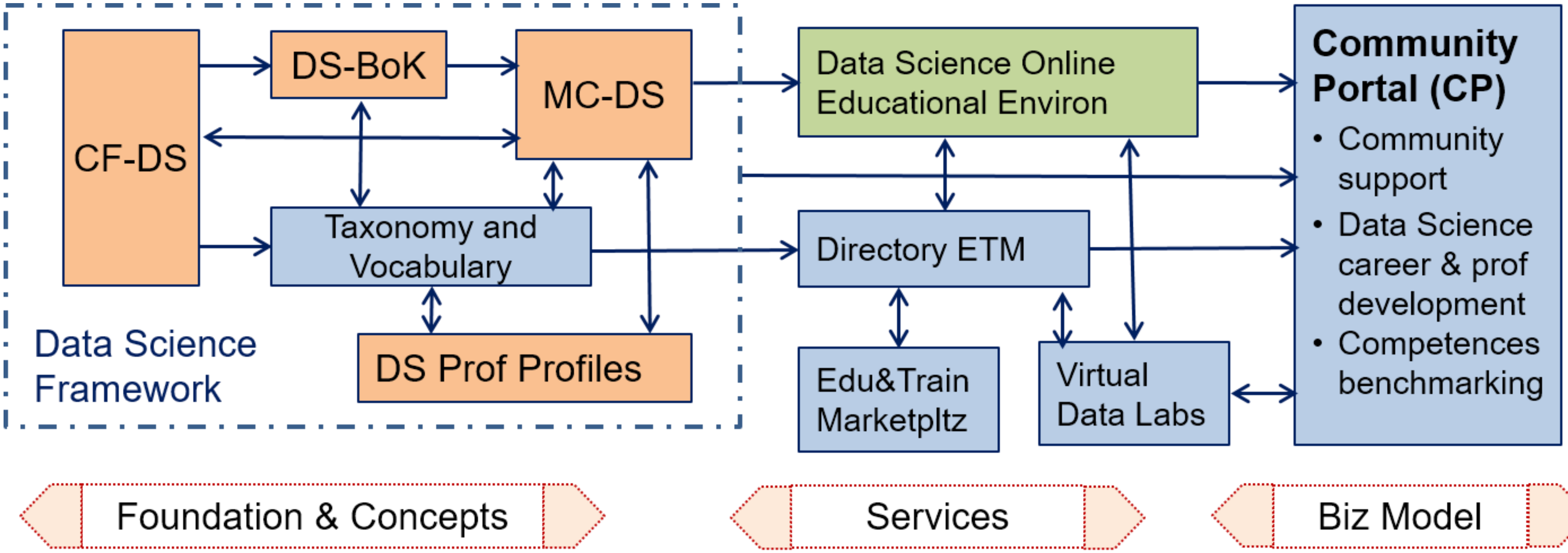


Data Science Applications Developer
Data Steward
Data Entry/ Support
Data Steward
Data Engineer, Database Developer
Data Collection
Data Ingest
Researcher (Scientific domain)
Data Facilities Operator
Data Source (Experiment, Data Driven Application)
Data Analysis
Data Scientist
Data Analyst/ Business Analyst
Data Science Applications Developer
Data Science Group Manager, Data Science Architect
Data Visualisation, Reporting, Storage
Results, Actionable Data
Data Steward
Data Science Researcher Business Analyst
Data Scientist
Data Scientist
Data Steward



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

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

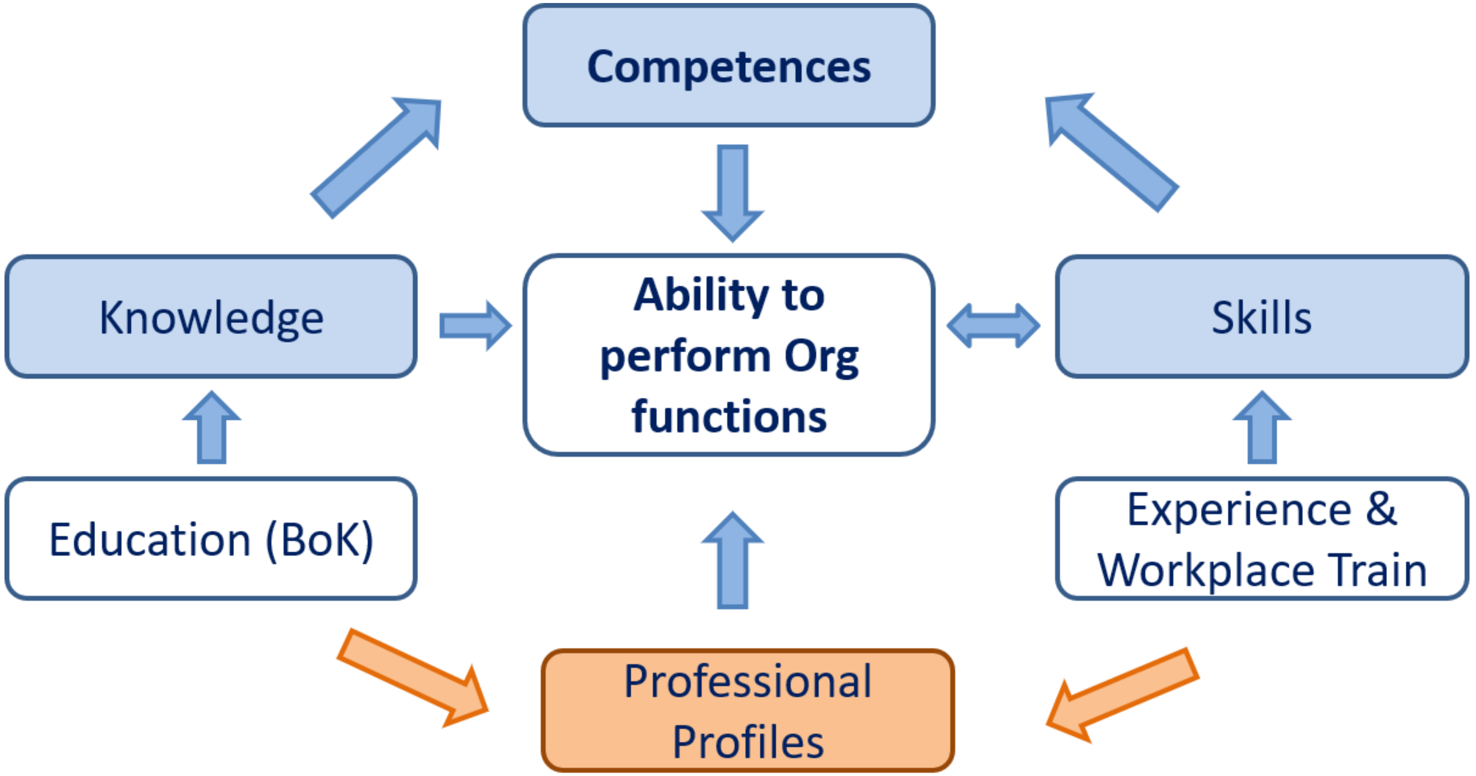
- ESDF development based on job market study, existing practices in academic, research and industry.
- Review and feedback from the ELG, expert community, domain experts.
- Input from the champion universities and community of practice.

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. Role of big technology companies in defining data-driven technology development
5. Correctly organise and manage data, make them accessible (adhering FAIR principles), education new profession of Data Stewards
6. Help managers to facilitate career dev for researchers and organise effective teams
7. Ensure skills and expertise sustain in organisation
8. Help research institutions to sustain in competition with industry and business in data science talent hunting

Competences Map to Knowledge and Skills

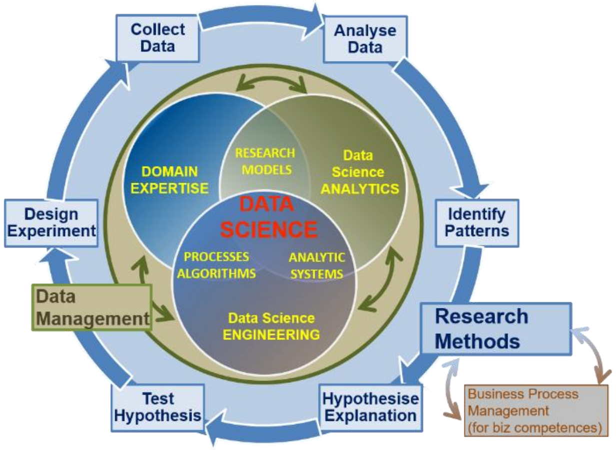
- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results



Data Scientist definition

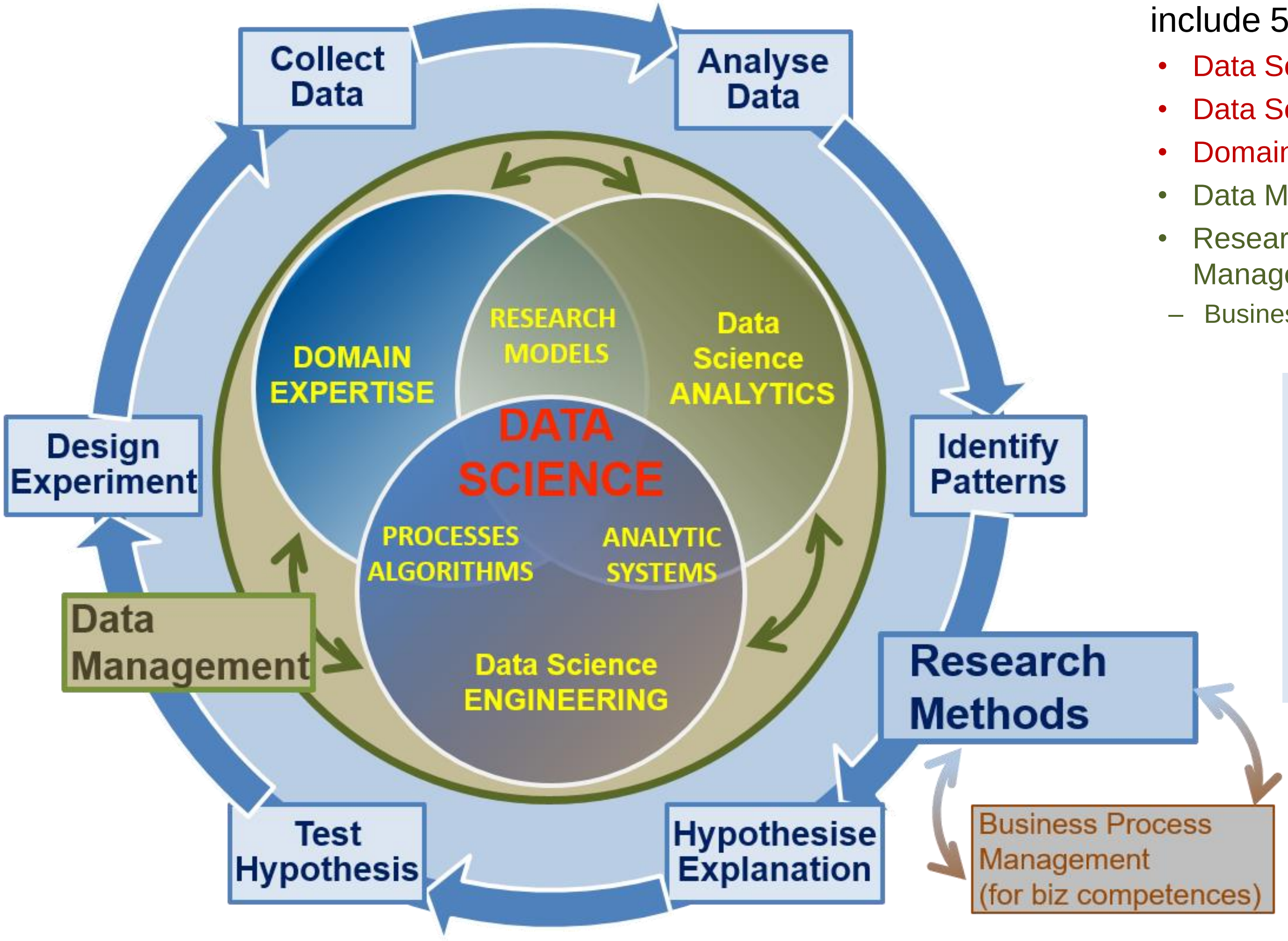
Based on the definitions by NIST SP1500 – 2015, extended by EDISON

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



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

Data Science Competence Groups - Research



Data Science Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Research Methods and Project Management
  - Business Process Management (biz)

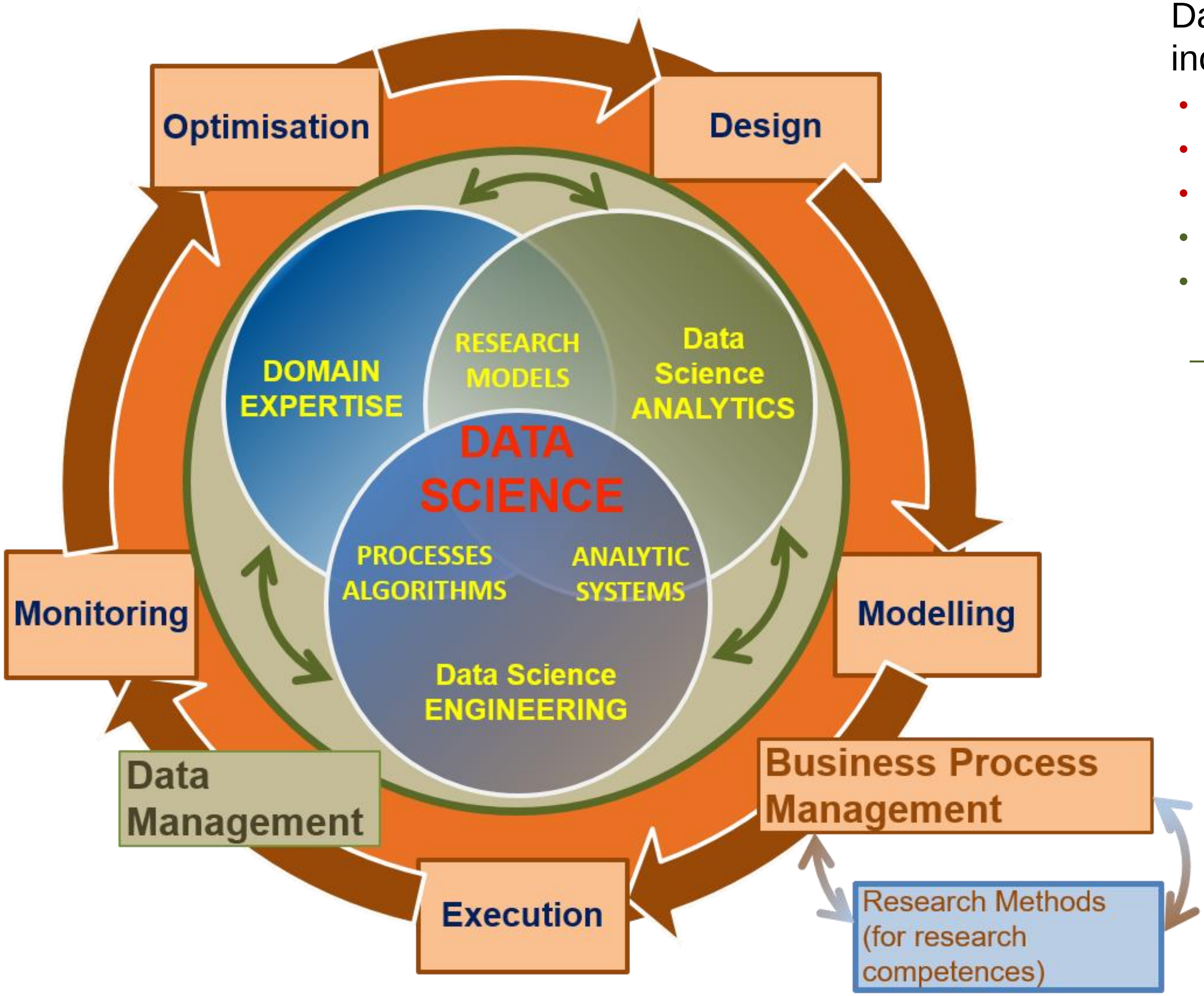
Scientific Methods

- 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

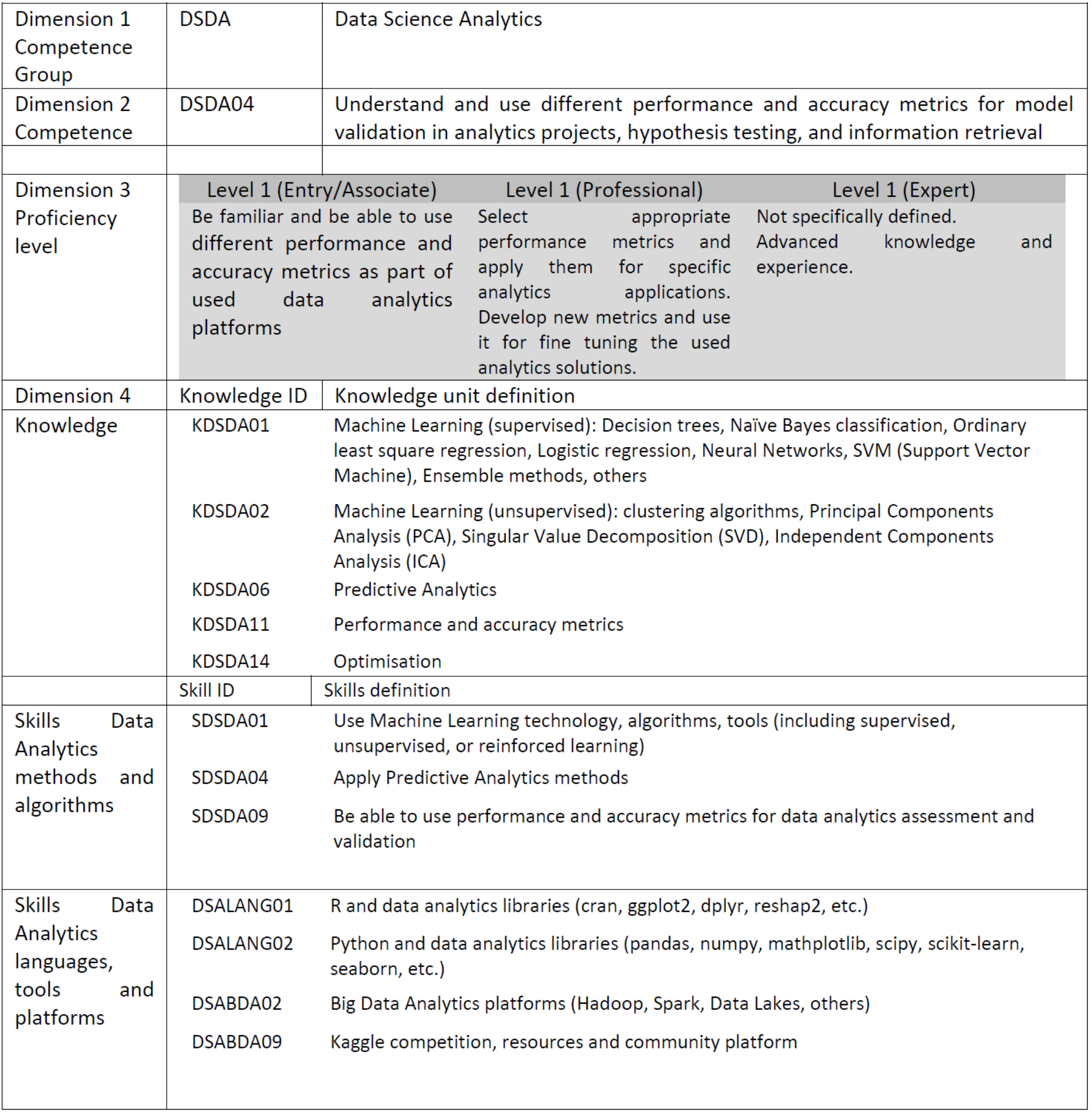
Skills Type A – Based on knowledge acquired

- **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

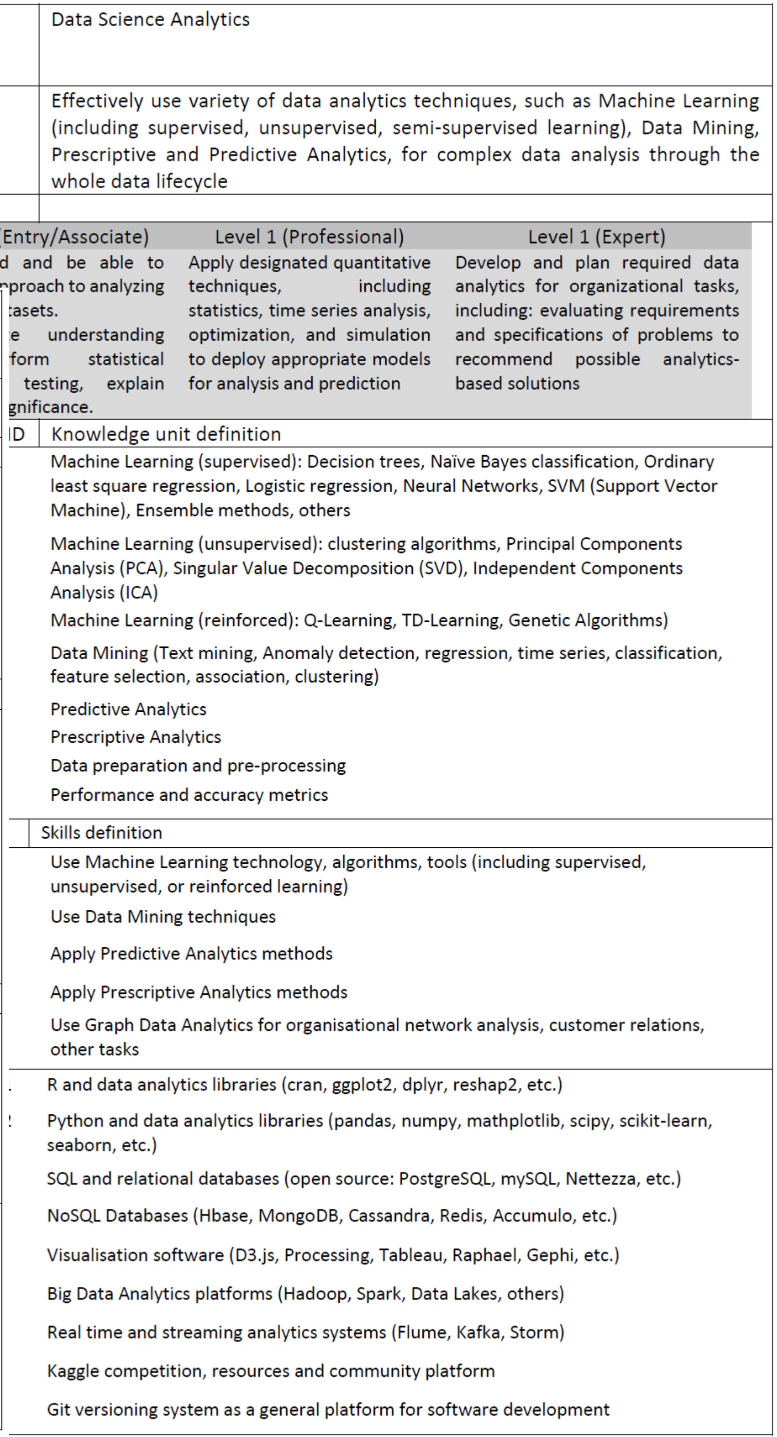
Skills Type B – Base on practical or workplace experience

- **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

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) Be familiar and be able to use different performance and accuracy metrics as part of used data analytics platforms	Level 1 (Professional) Select appropriate performance metrics and apply them for specific analytics applications. Develop new metrics and use it for fine tuning the used analytics solutions. Level 1 (Expert) 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



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) Understand and be able to select an approach to analyzing ...tasets. ...e understanding ...form statistical ... testing, explain ...gnificance.	Level 1 (Professional) Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction Level 1 (Expert) Develop and plan required data analytics for organizational tasks, including: evaluating requirements and specifications of problems to recommend possible analytics-based solutions
	ID	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

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 net**works and communities
13. **Story Telling**: Deliver actionable result of your analysis
14. **Attitude**: Creativity, curiosity (willingness to challenge status quo), commitment in finding new knowledge and progress to completion
15. **Ethics and responsible use** of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)

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

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

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

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

21st Century Skills (DARE & BHEF & EDISON)

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

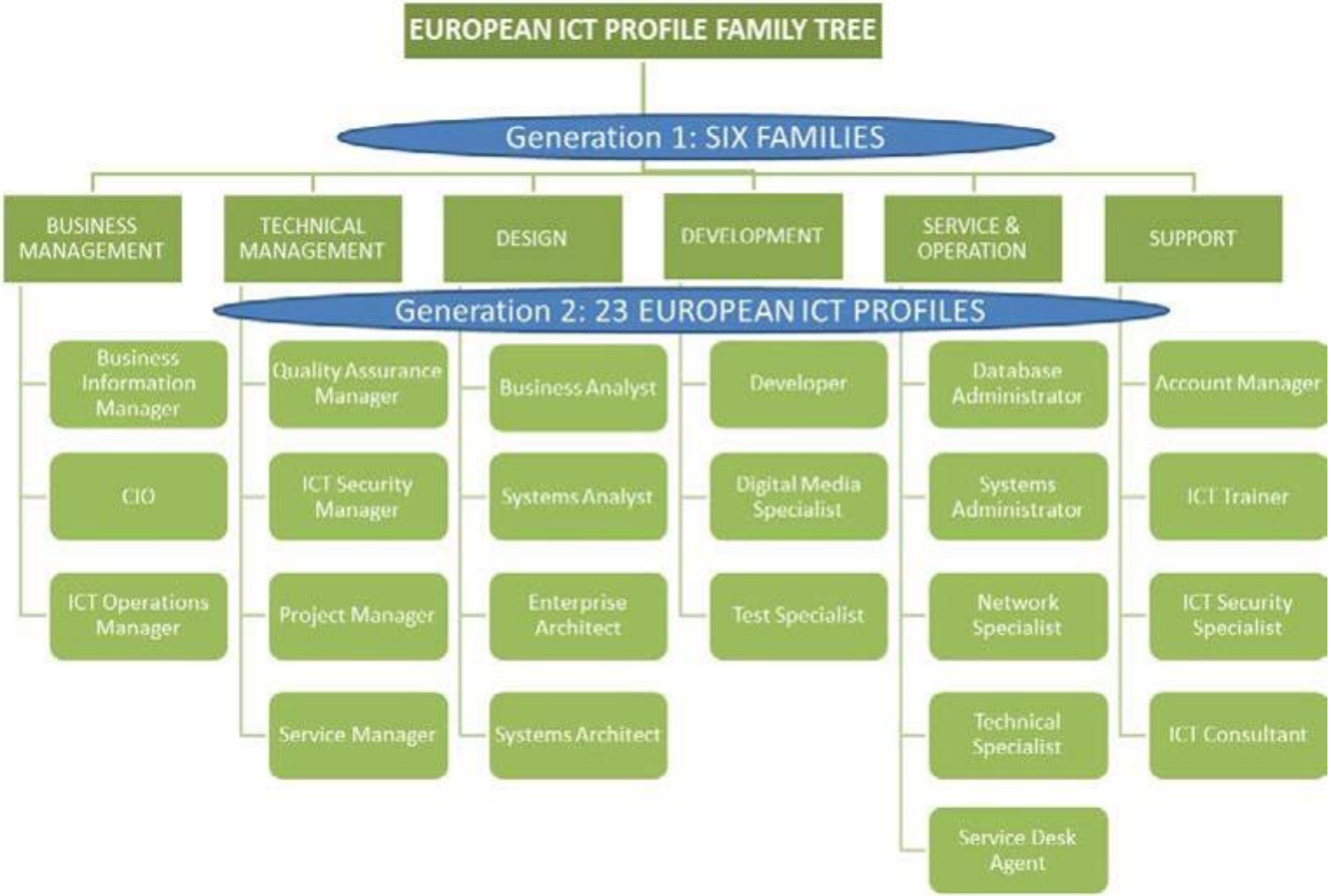
Practical Application of the CF-DS

- Basis for the definition of the Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS)
  - CF-DS => Learning Outcomes (MC-DS) => Knowledge Areas (DS-BoK)
  - CF-DS => Data Science taxonomy of scientific subjects and vocabulary
- Data Science professional profiles definition
  - Extend existing EU standards and occupations taxonomies: e-CFv3.0, ESCO, others
- Professional competence ***benchmarking***
  - For customizable training and career development
  - Including CV or organisational profiles matching
- ***Professional certification***
  - In combination with DS-BoK professional competences benchmarking
- Vacancy construction tool for job advertisement (for HR)
  - Using controlled vocabulary and Data Science Taxonomy
  - Candidates' CV assessment

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



EUROPEAN ICT PROFILE FAMILY TREE

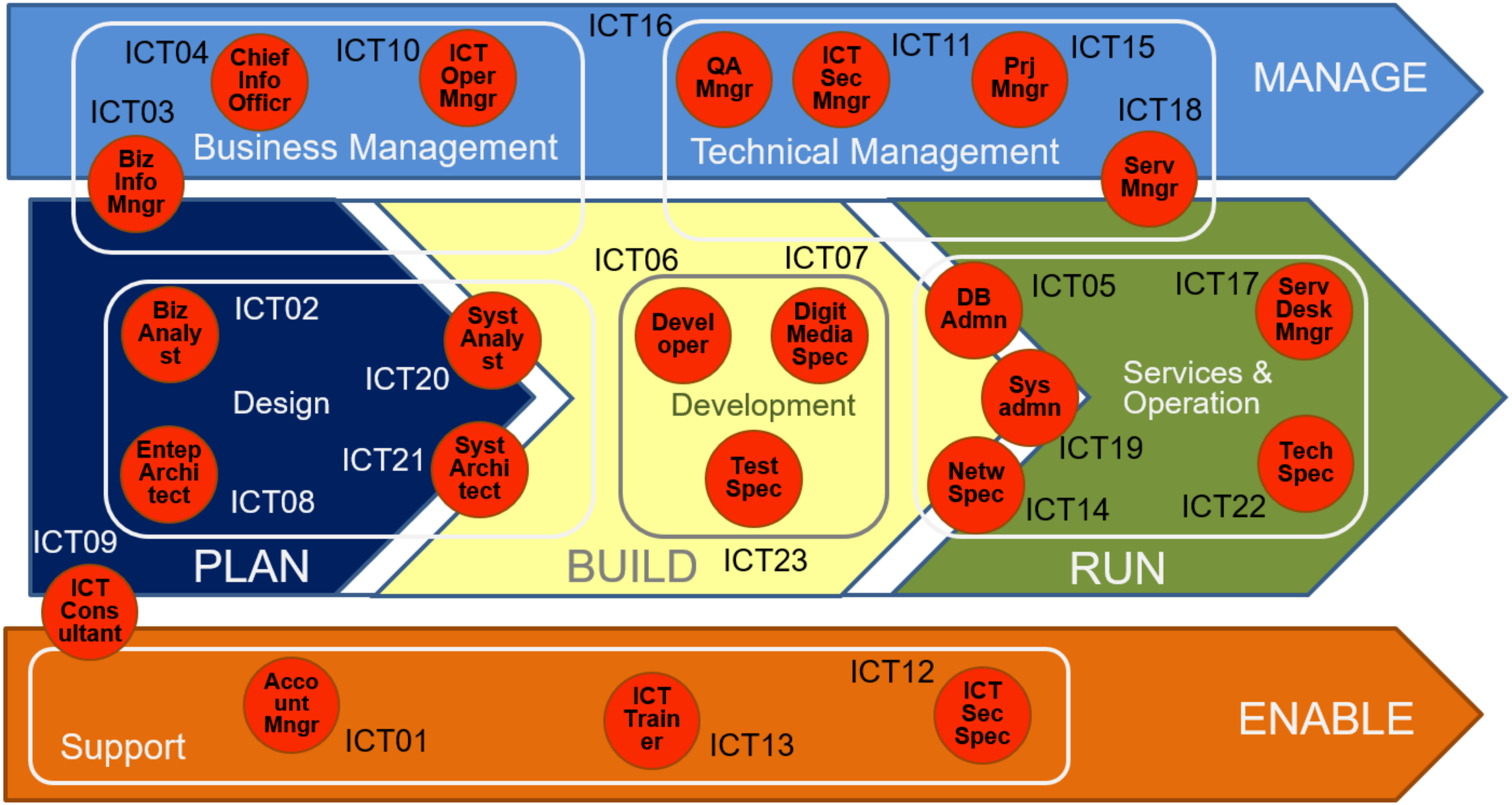
Generation 1: SIX FAMILIES

Generation 2: 23 EUROPEAN ICT PROFILES

- BUSINESS MANAGEMENT
  - Business Information Manager
  - CIO
  - ICT Operations Manager
- TECHNICAL MANAGEMENT
  - Quality Assurance Manager
  - ICT Security Manager
  - Project Manager
  - Service Manager
- DESIGN
  - Business Analyst
  - Systems Analyst
  - Enterprise Architect
  - Systems Architect
- DEVELOPMENT
  - Developer
  - Digital Media Specialist
  - Test Specialist
- SERVICE & OPERATION
  - Database Administrator
  - Systems Administrator
  - Network Specialist
  - Technical Specialist
  - Service Desk Agent
- SUPPORT
  - Account Manager
  - ICT Trainer
  - ICT Security Specialist
  - ICT Consultant

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

CWA Professional Profiles and Organisational Workflow



Data Science Professions Family

Managers: Chief Data Officer (CDO), Data Science (group/dept) manager, Data Science infrastructure manager, Research Infrastructure manager

DSP 01, DSP 02, DSP 03

Professionals: Data Scientist, Data Science Researcher, Data Science Architect, Data Science (applications) programmer/engineer, Data Analyst, Business Analyst, etc.

DSP 04, DSP 05, DSP 06, DSP 07, DSP 08, DSP 09

Professional (database): Large scale (cloud) database designers and administrators, scientific database designers and administrators

DSP 14, DSP 15, DSP 16

Professional (data handling/management): Data Stewards, Digital Data Curator, Data Librarians, Data Archivists

DSP 10, DSP 11, DSP 12, DSP 13

Technicians and associate professionals: Big Data facilities operators, scientific database/infrastructure operators

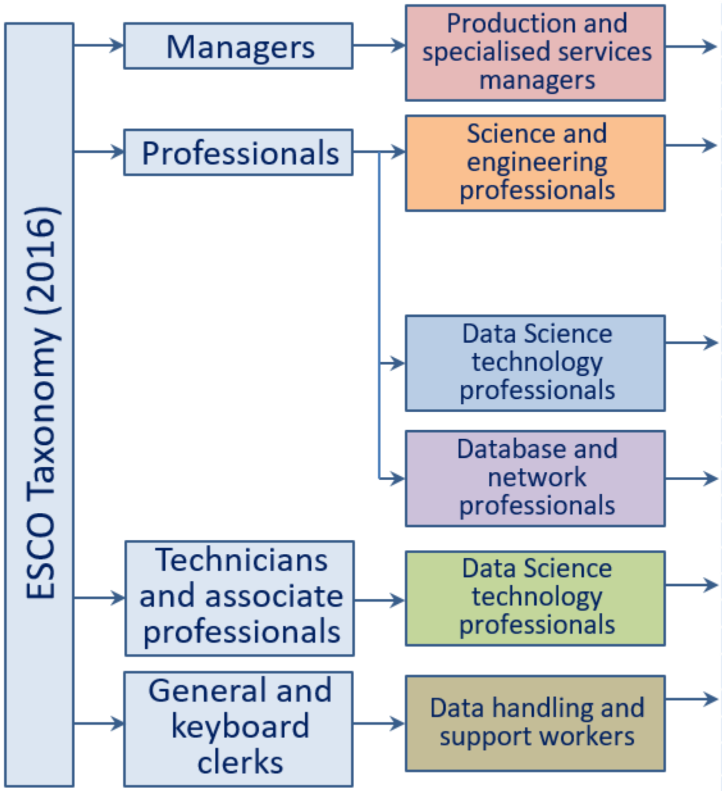
DSP 17, DSP 18, DSP 19

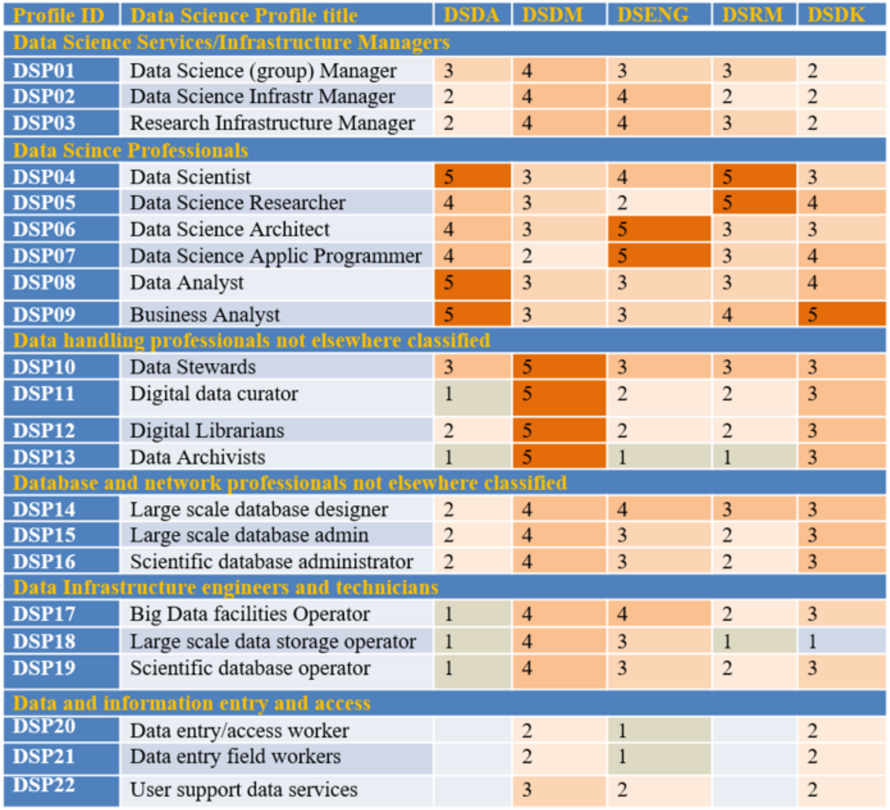
Support workers and data handling clerks: User support workers, data entry clerks, data entry field workers

DSP 20, DSP 21, DSP 22

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

DSP Profiles mapping to ESCO Taxonomy High Level Groups

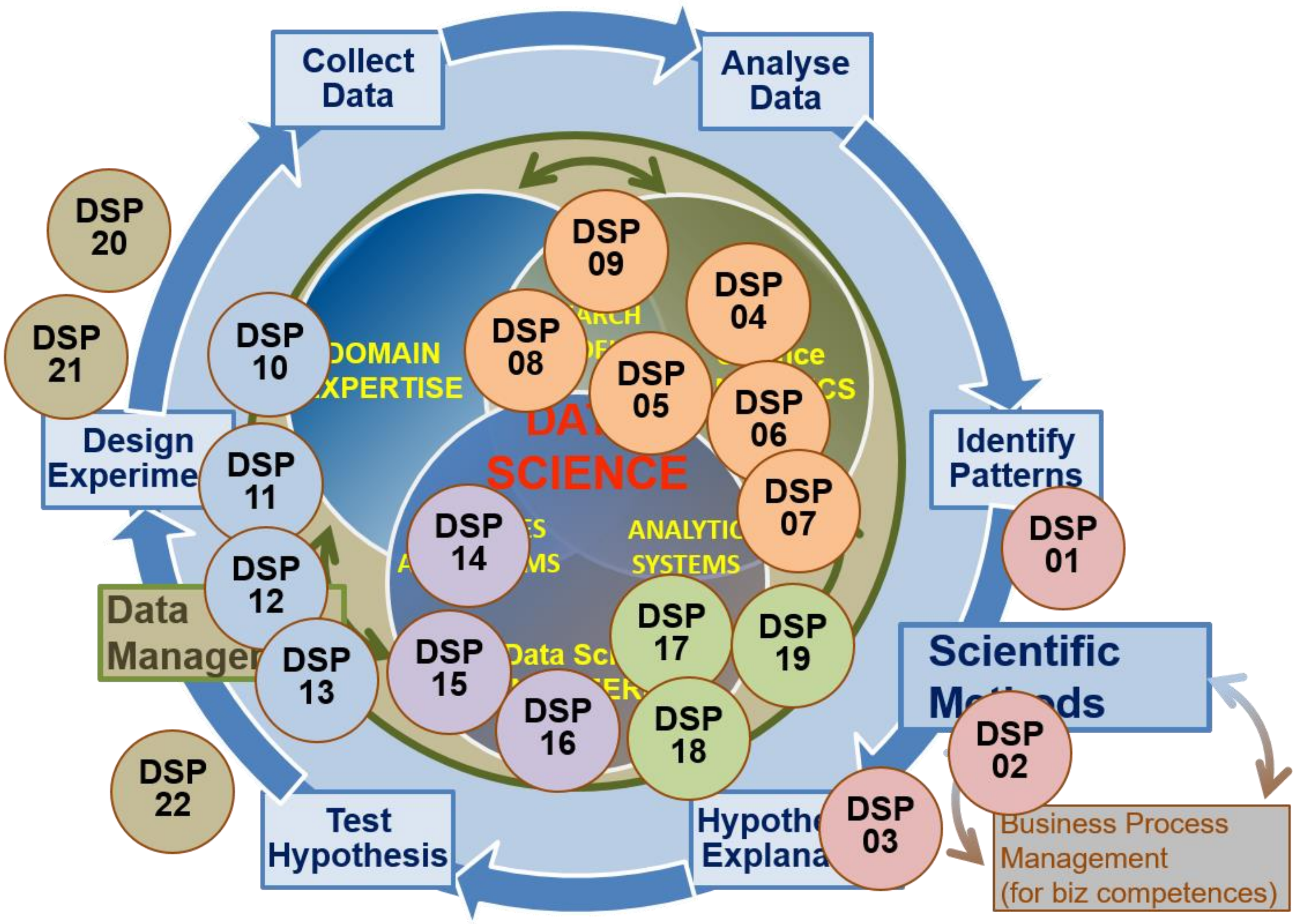

ESCO Taxonomy (2016)
Managers
Professionals
Technicians and associate professionals
General and keyboard clerks
Production and specialised services managers
Science and engineering professionals
Data Science technology professionals
Database and network professionals
Data Science technology professionals
Data handling and support workers



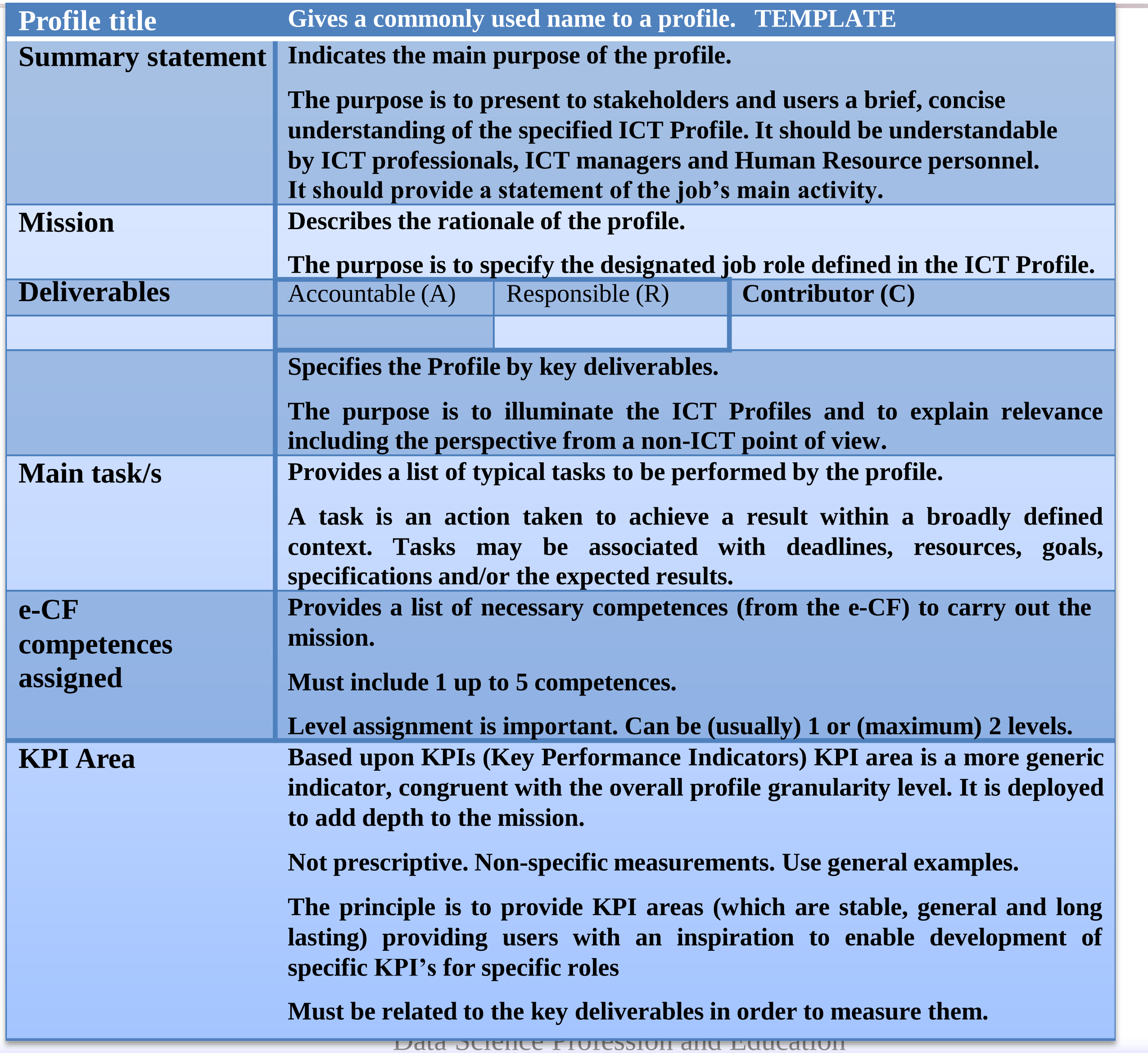
Profile ID	Data Science Profile title	DSDA	DSDM	DSENG	DSRM	DSDK
Data Science Services/Infrastructure Managers						
DSP01	Data Science (group) Manager	3	4	3	3	2
DSP02	Data Science Infrastr Manager	2	4	4	2	2
DSP03	Research Infrastructure Manager	2	4	4	3	2
Data Scince Professionals						
DSP04	Data Scientist	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 - in compliance with CWA 16458 (2012)



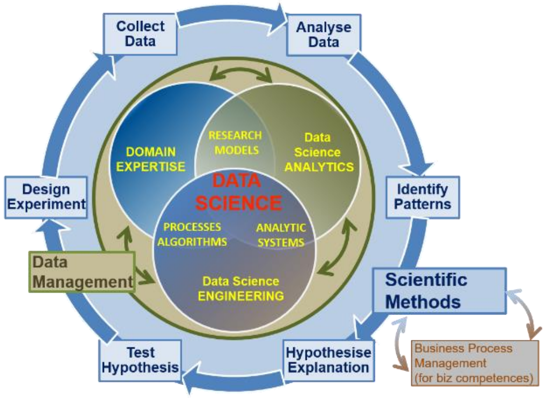
Profile title	Gives a commonly used name to a profile. TEMPLATE		
Summary statement	Indicates the main purpose of the profile. The purpose is to present to stakeholders and users a brief, concise understanding of the specified ICT Profile. It should be understandable by ICT professionals, ICT managers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

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

MC-DS: 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.

MC-DS: 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

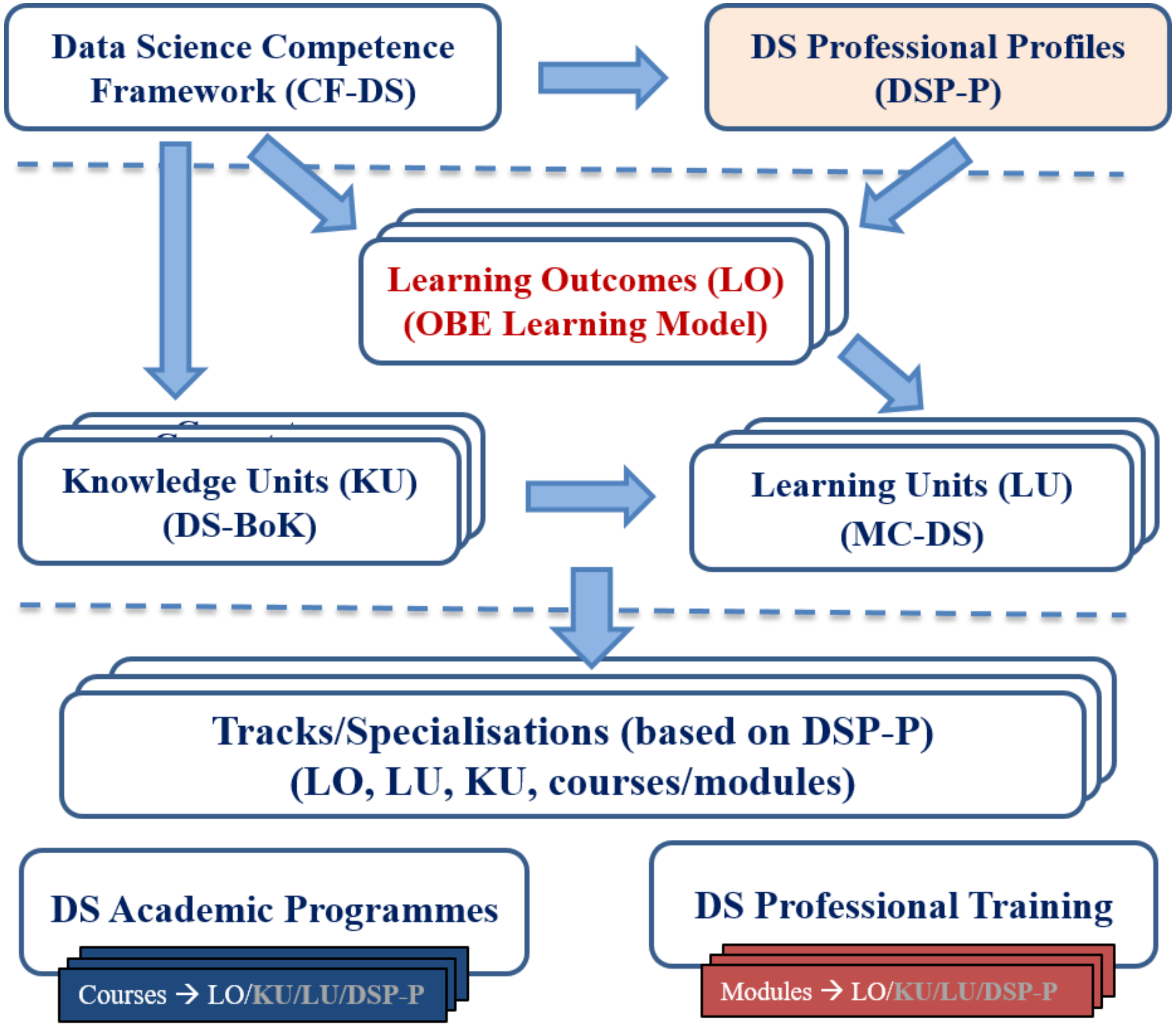
- (1) Data Governance
- (2) Data Architecture
- (3) Data Modelling and Design
- (4) Data Storage and Operations
- *(5) Data Security*
- (6) Data Integration and Interoperability
- *(7) Documents and Content*
- (8) Reference and Master Data
- (9) Data Warehousing and Business Intelligence
- ***(10) Metadata***
- (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



Data Science Competence Framework (CF-DS)

DS Professional Profiles (DSP-P)

Learning Outcomes (LO) (OBE Learning Model)

Knowledge Units (KU) (DS-BoK)

Learning Units (LU) (MC-DS)

Tracks/Specialisations (based on DSP-P) (LO, LU, KU, courses/modules)

DS Academic Programmes

Courses → LO/KU/LU/DSP-P

DS Professional Training

Modules → LO/KU/LU/DSP-P

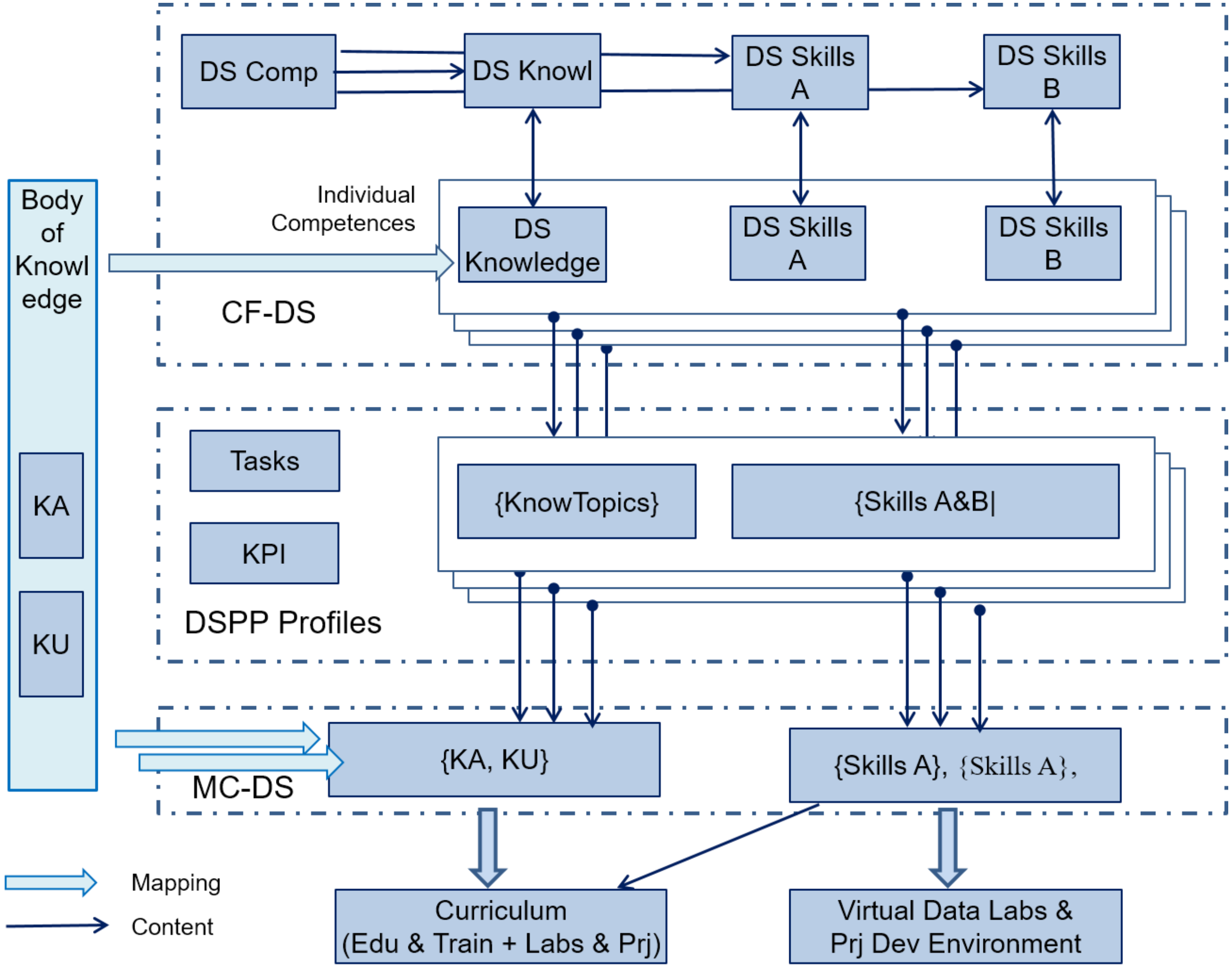
From Competences and DSP Profiles

to Learning Outcomes (LO) and

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

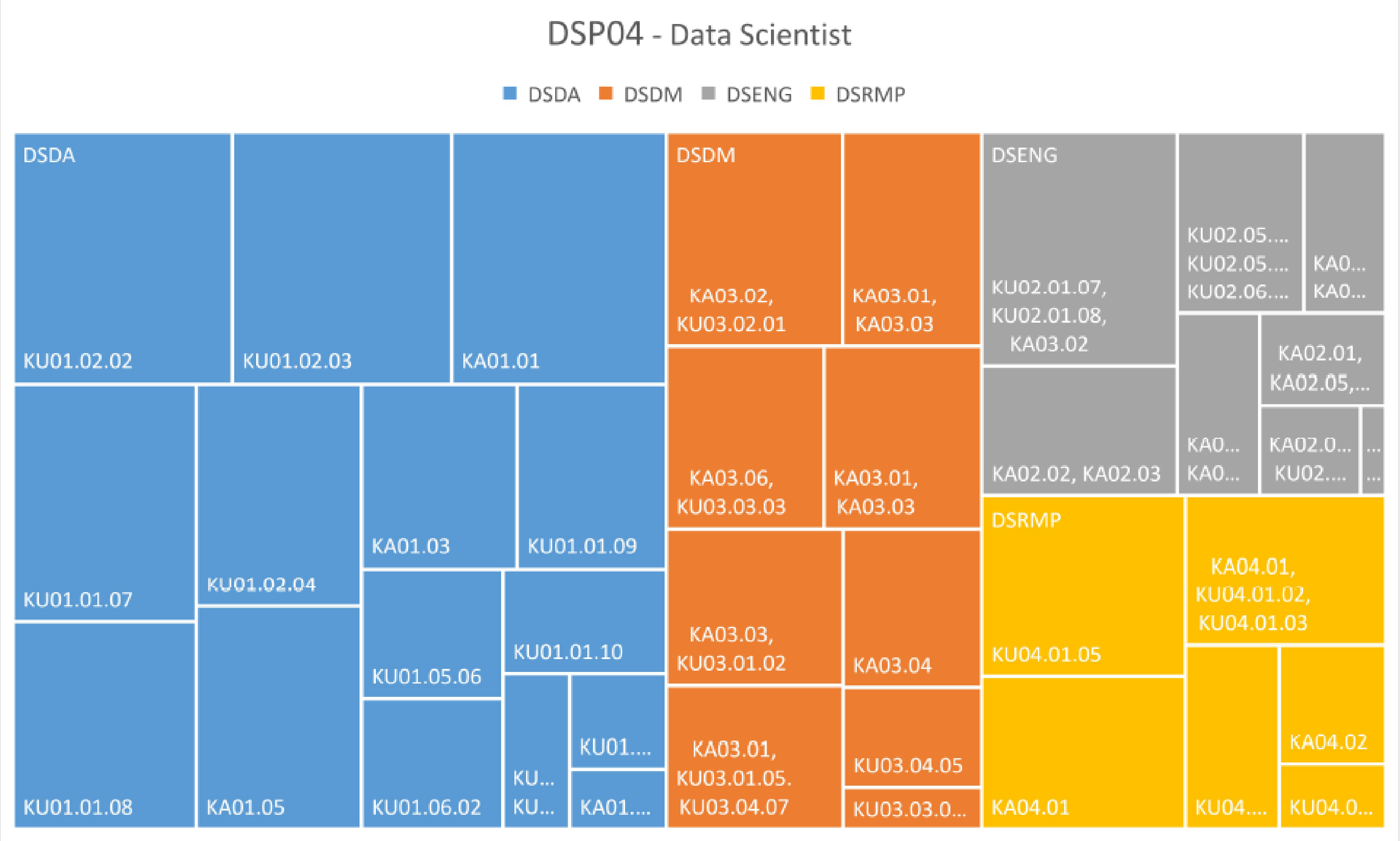
- EDSF allow for customized educational courses and training modules design

EDSF Data Model and API

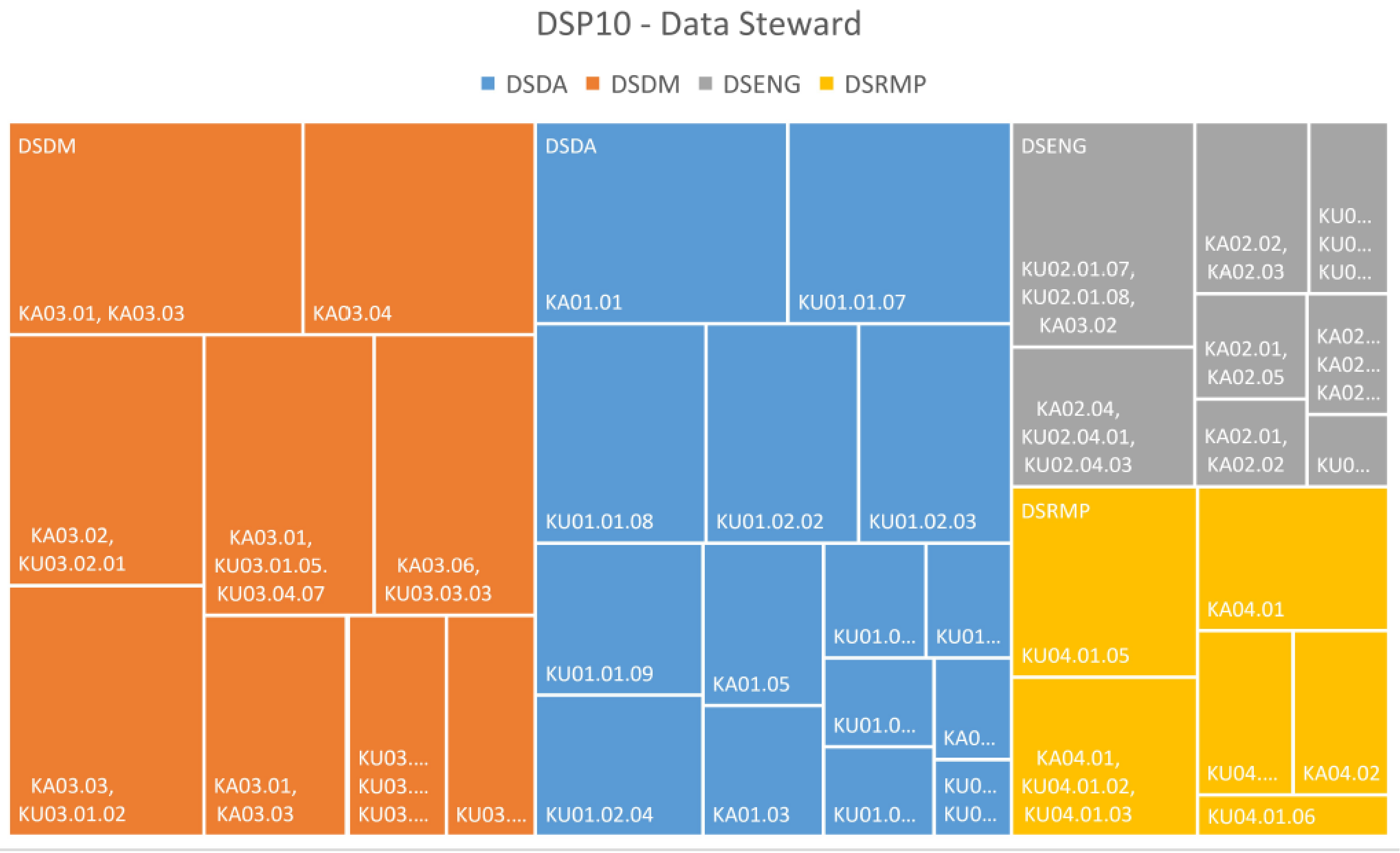


- EDSF API provides access to all EDSF functionality

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:

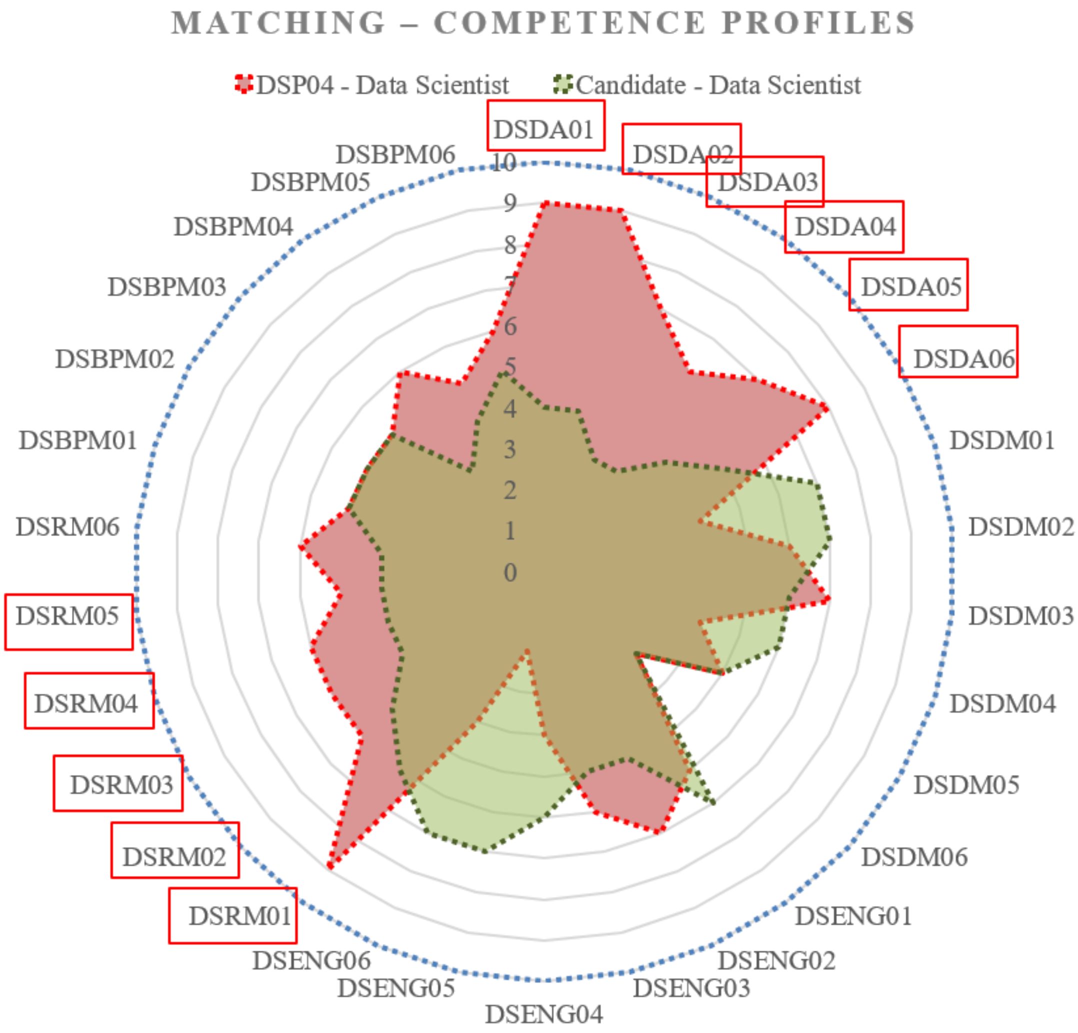
- Python (or R) and corresponding data analytics libraries
- NoSQL and SQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, MS SQL, My SQL, PostgreSQL, etc.)
- Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
- Real time and streaming analytics systems (Flume, Kafka, Storm)
- Kaggle 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 adjustment

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)

Individual Competences Benchmarking

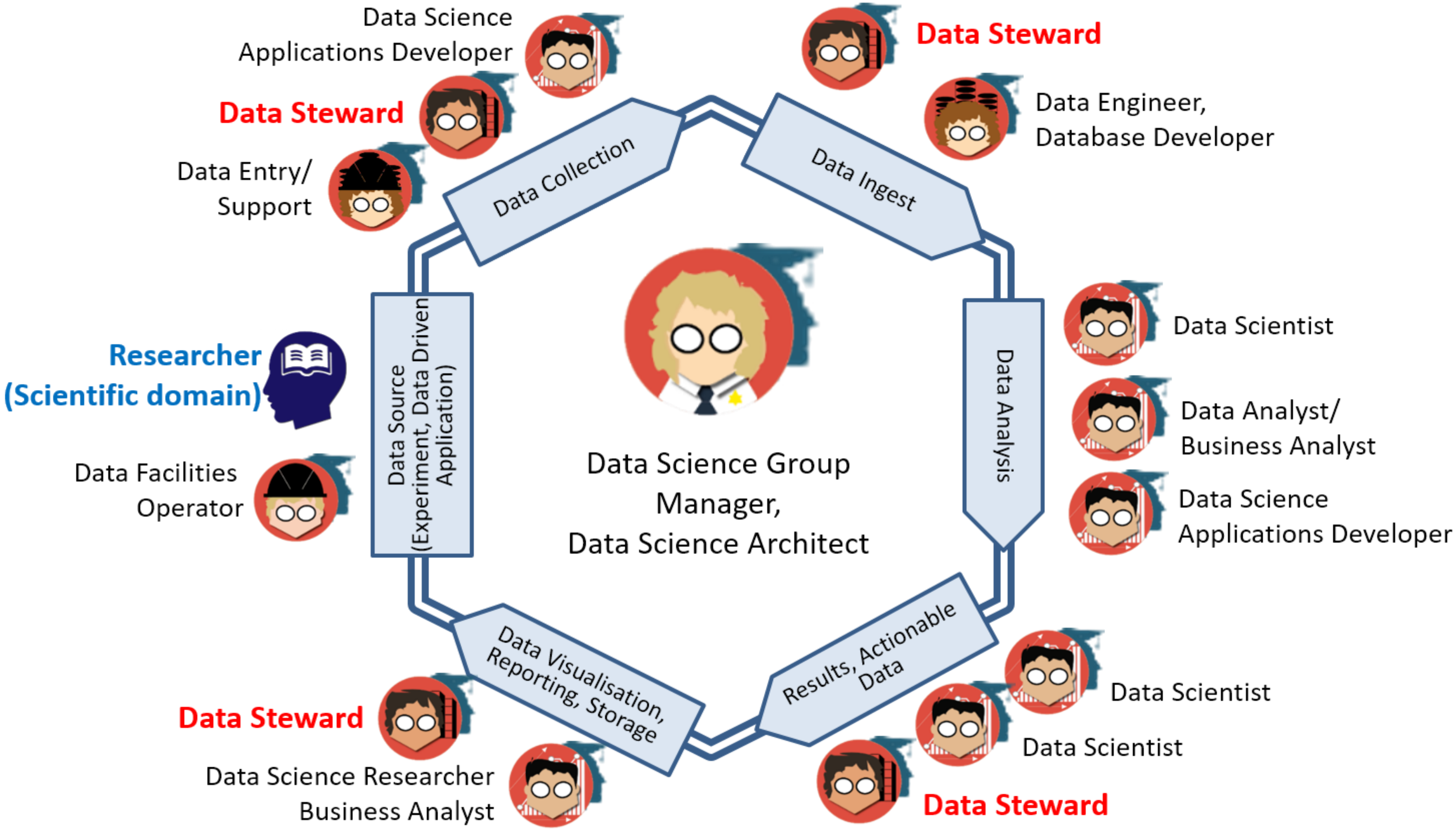


Individual Education/Training Path based on Competence benchmarking

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

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

Building a Data Science Team



Data Science or Data Management Group/Department: Organisational structure and staffing - EXAMPLE

Data Science or Data Management Group/Department

>> Reporting to CDO/CTO/CEO

- Providing cross-organizational services

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

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

Growing role and demand for Data Stewards and data stewardship

Data Stewards – A rising new role in Data Science ecosystem

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

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

A. Use cases for data management and stewardship

- Preserving the Scientific Record

B. Data Management elements (organisational and individual)

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

Collaboration with the Research Data Alliance (RDA) on developing model curriculum on Research Data Literacy:

- Modular, Customisable, Localised, Open Access
- Supported by the network of trainers via resource swap board

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

D. Open Science and Open Data (Definition, Standards, Open Data use and reuse, open government data)

- Research data and open access
- Repository and self- archiving services
- ORCID identifier for data
- Stakeholders and roles: engineer, librarian, researcher
- Open Data services: ORCID.org, Altmetric Doughnut, Zenodo

E. Hands on:

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

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.

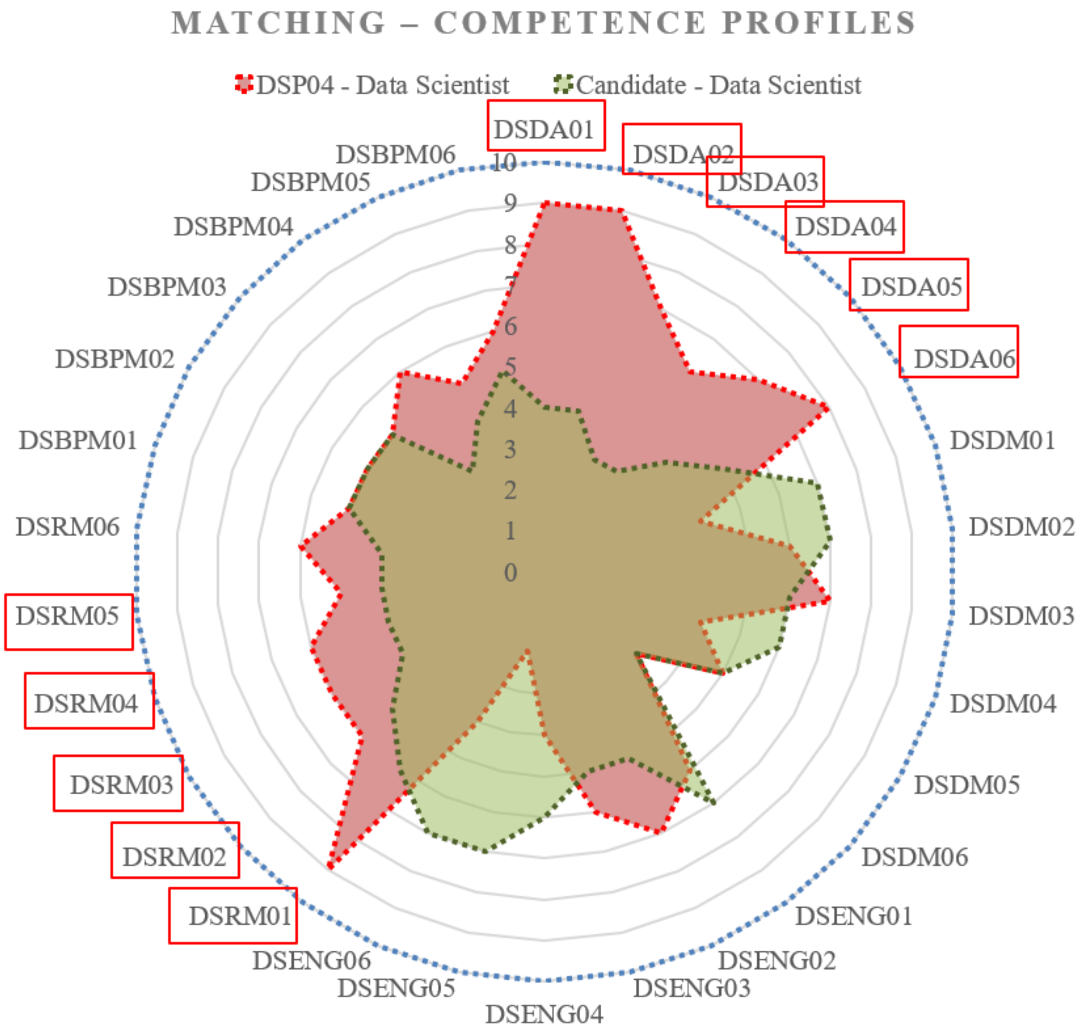
Discussion: How to become a Data Scientist

- 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

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

Individual Competences Benchmarking

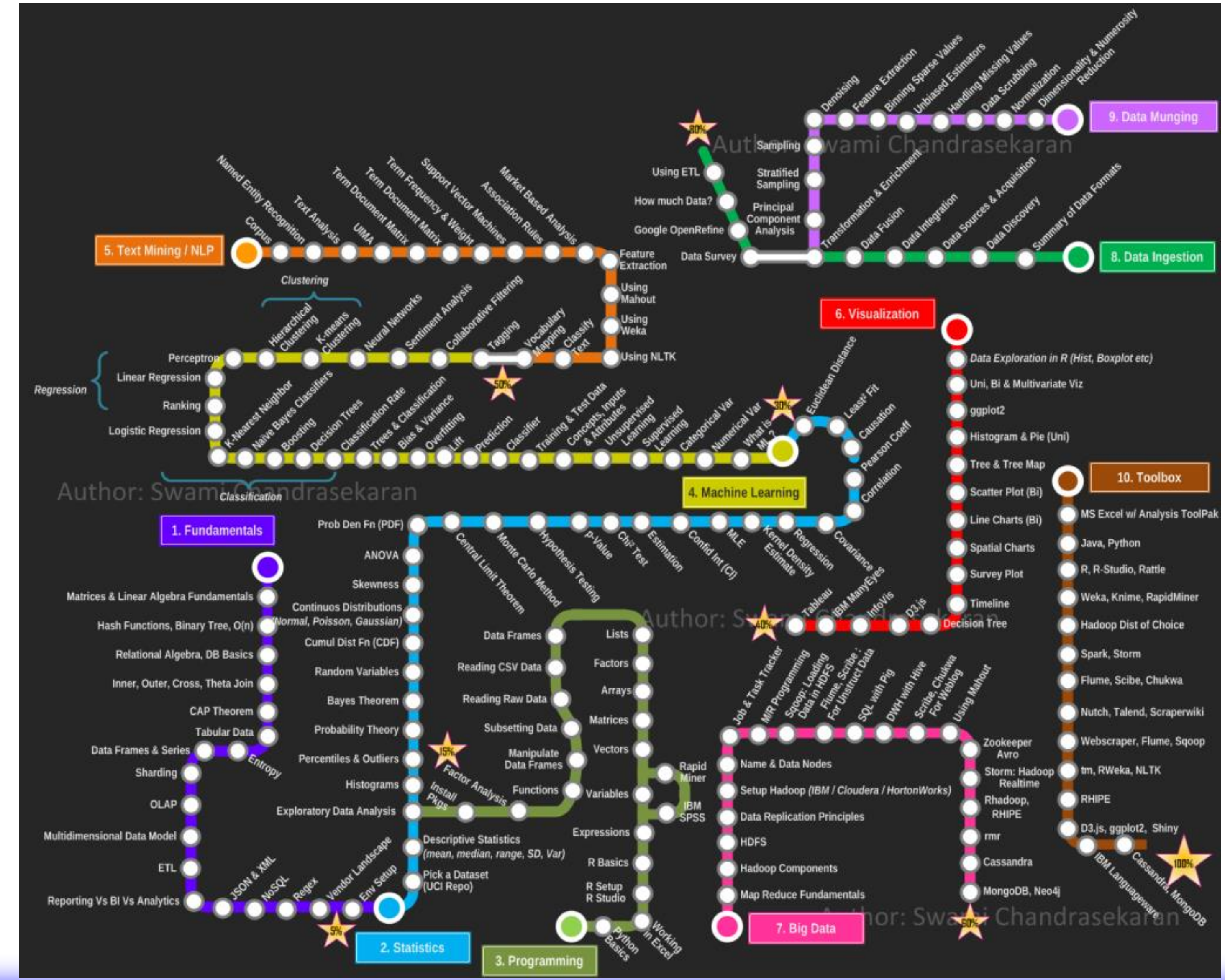


Individual Education/Training Path based on Competence benchmarking

- 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.

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

Online Educational and training resources

- LinkedIn Education
- Microsoft Virtual Academy
- (IBM – in transition)
- DataCamp
- Coursera, Udacity
- Certification and training PMI, DAMA, IIBA

Open Data and Educational Datasets

- Amazon Web Services (AWS)
- Google
- Microsoft Azure
- Kaggle
- KDNuggets
- Emerging - https://www.datasciencepro.eu/

Questions and discussion

Links to EDISON Resources

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

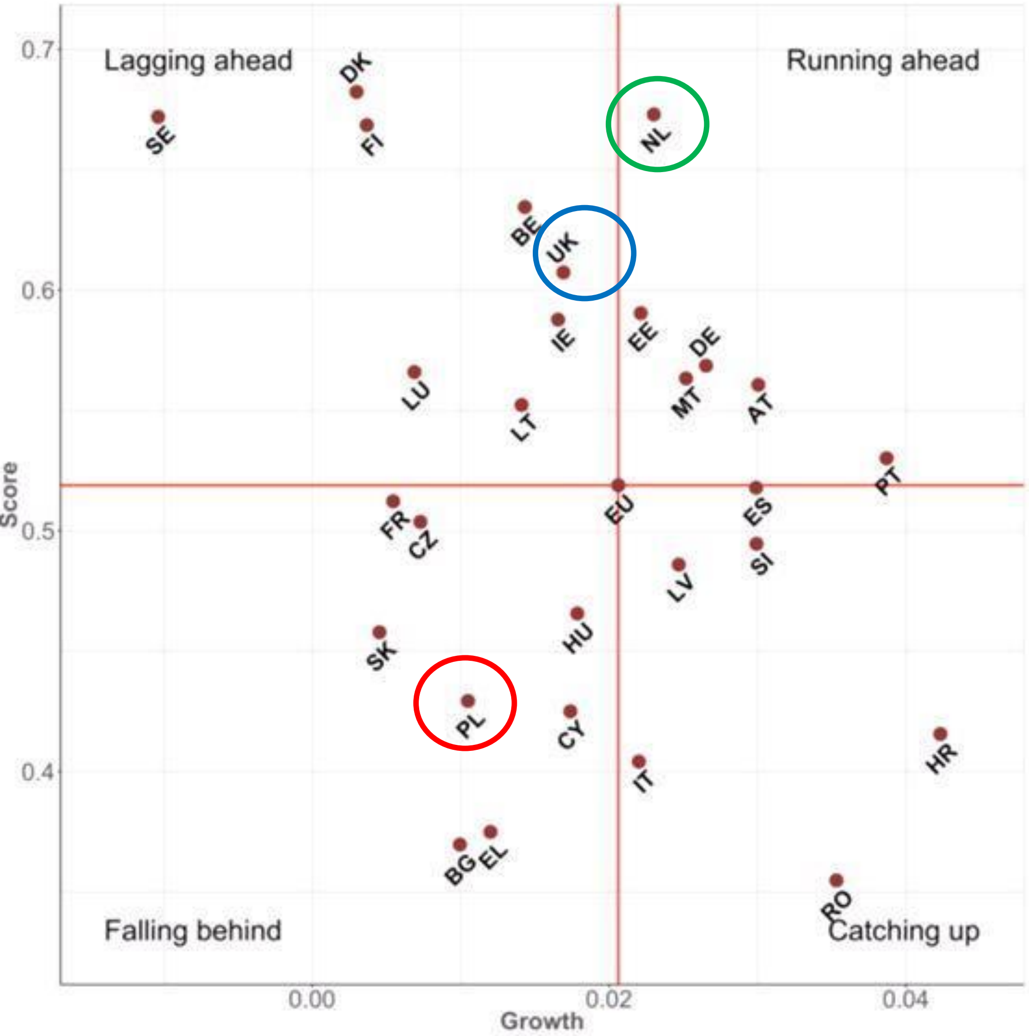
Other related links

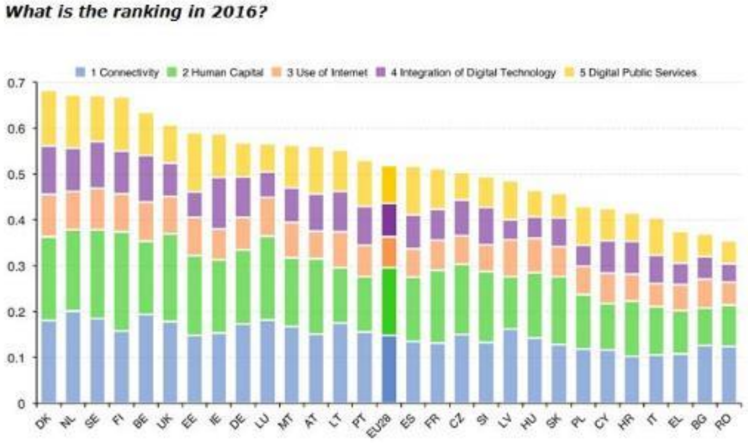
- Amsterdam School of Data Science
  - https://www.schoolofdatascience.amsterdam/
  - https://www.schoolofdatascience.amsterdam/education/

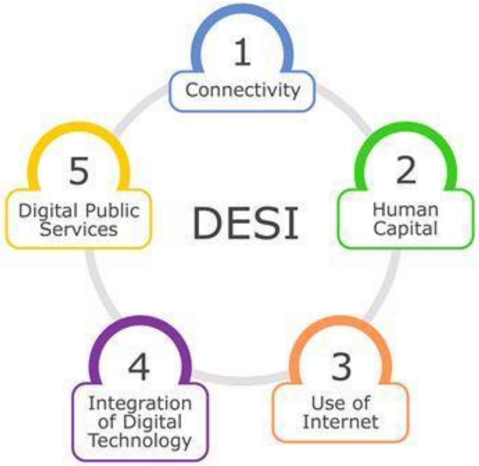
- Research Data Alliance interest Group on Education and Training on Handling of Research Data (IG-ETHRD)
  - https://www.rd-alliance.org/groups/education-and-training-handling-research-data.html
- Final Report on European Data Market Study by IDC (Feb 2017)
  - https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring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

Additional materials

Digital Economy and Society Index (EU 2015-2016)






DESI

1 Connectivity

2 Human Capital

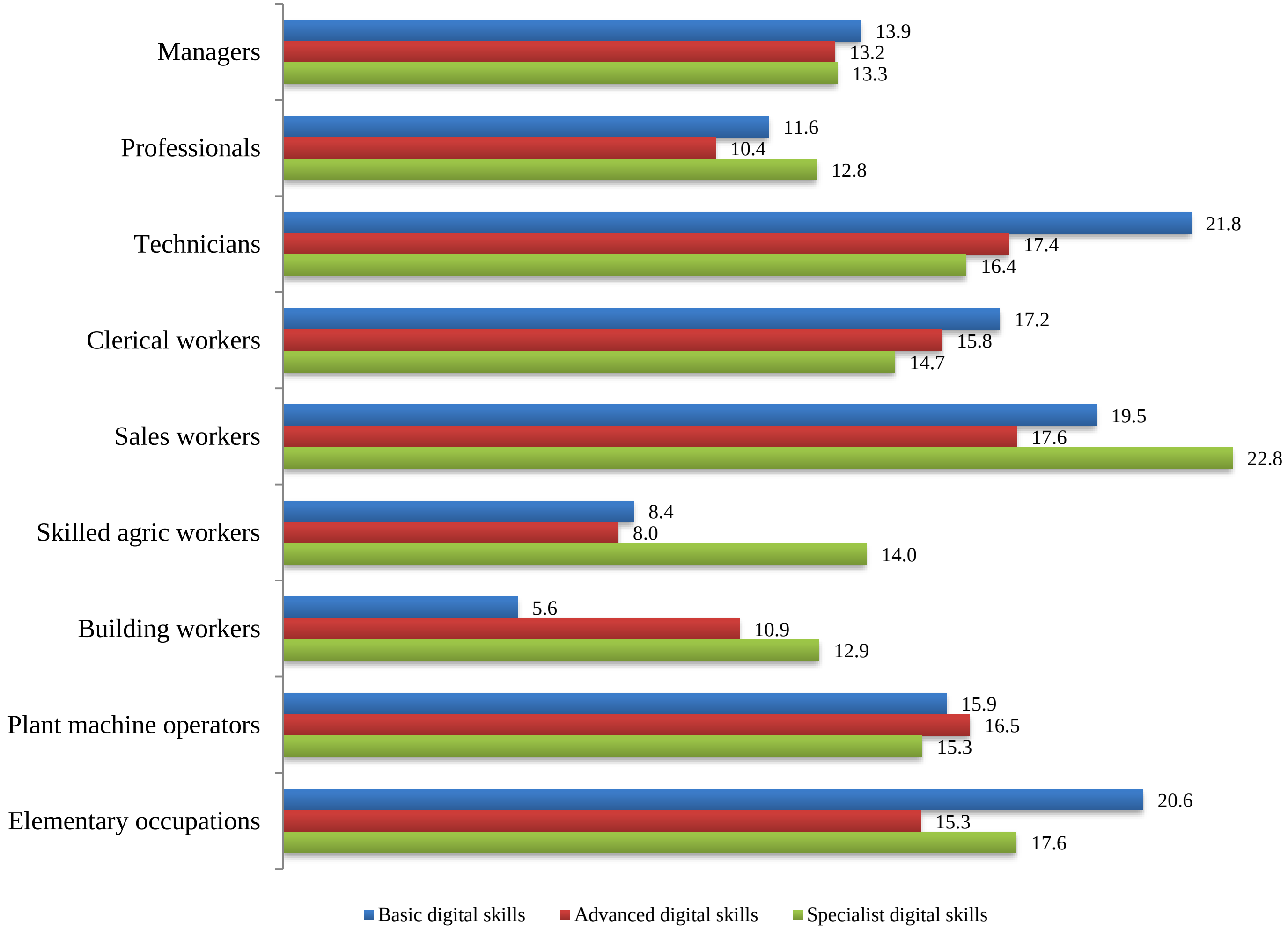
3 Use of Internet

4 Integration of Digital Technology

5 Digital Public Services

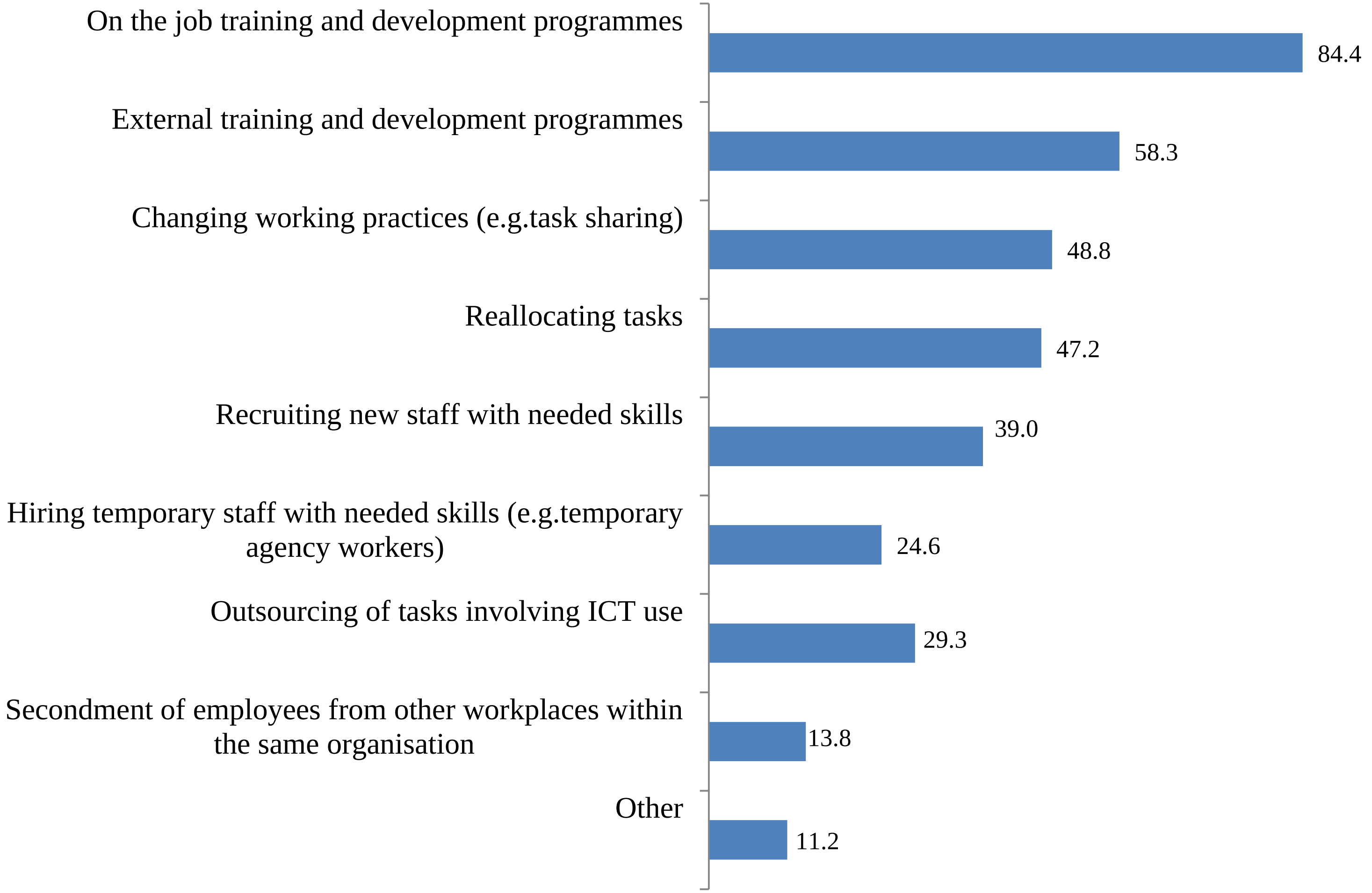
[ref] http://europa.eu/rapid/press-release_MEMO-16-385_en.htm

Digital skills gaps density by occupation and type of digital skills, EU28 (%)



ICT for work: Digital skills in the workplace, Digital Single Market, Reports and studies, May 2017
https://ec.europa.eu/digital-single-market/en/news/ict-work-digital-skills-workplace

Workplaces reporting having taken action to tackle digital skill gaps by type of action undertaken, EU28 (% of workplaces with digital skill gaps which undertook actions)

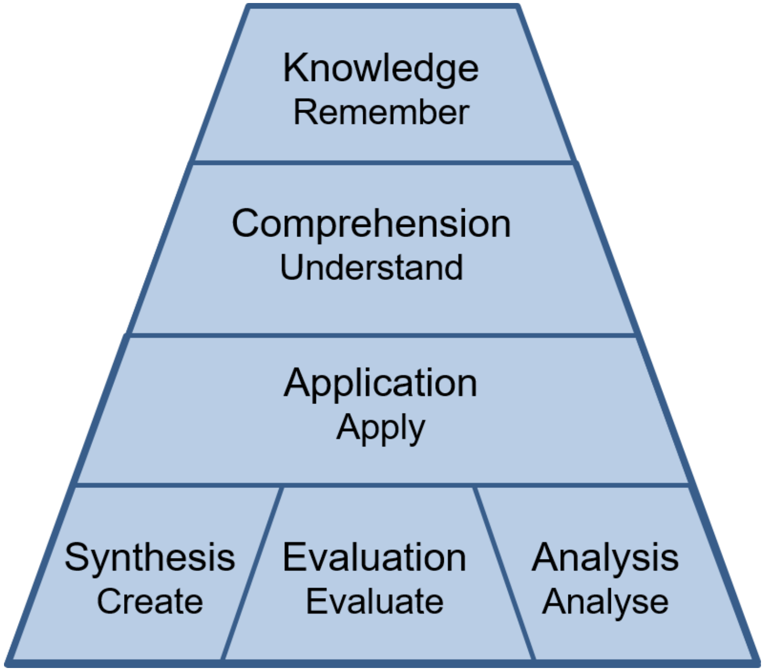


ICT for work: Digital skills in the workplace, Digital Single Market, Reports and studies, May 2017
https://ec.europa.eu/digital-single-market/en/news/ict-work-digital-skills-workplace

EDSF Recognition, Endorsement and Implementation

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

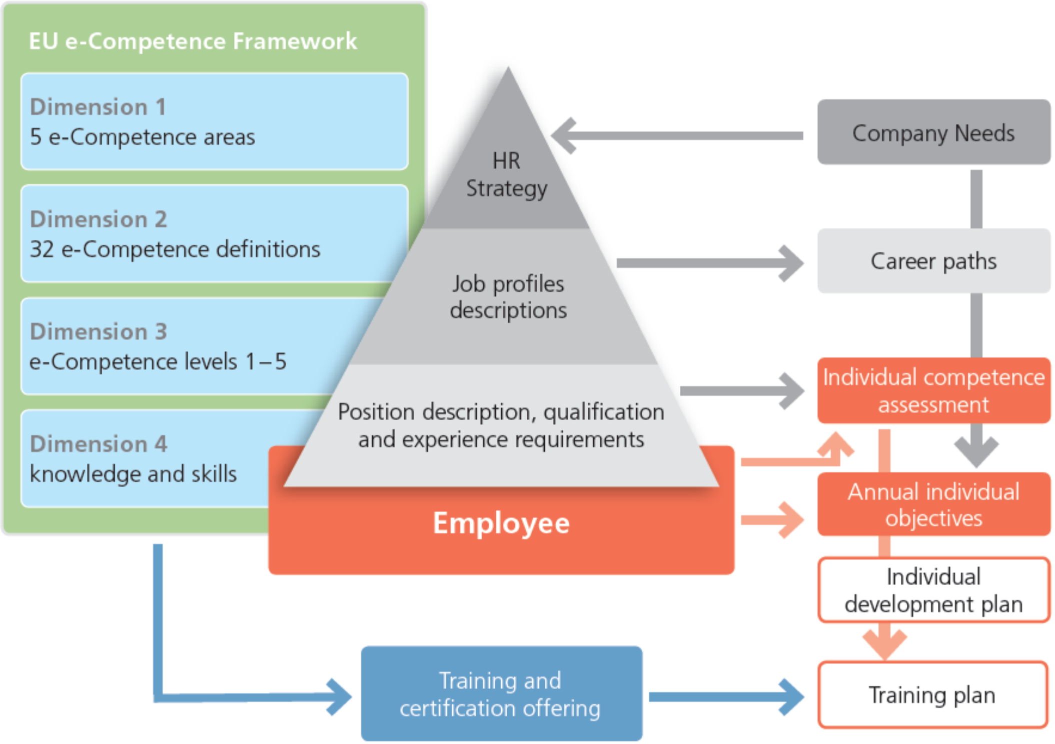
Bloom’s Taxonomy and Knowledge Levels for MC-DS



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

e-CFv3.0 structure and 4-dimensional model

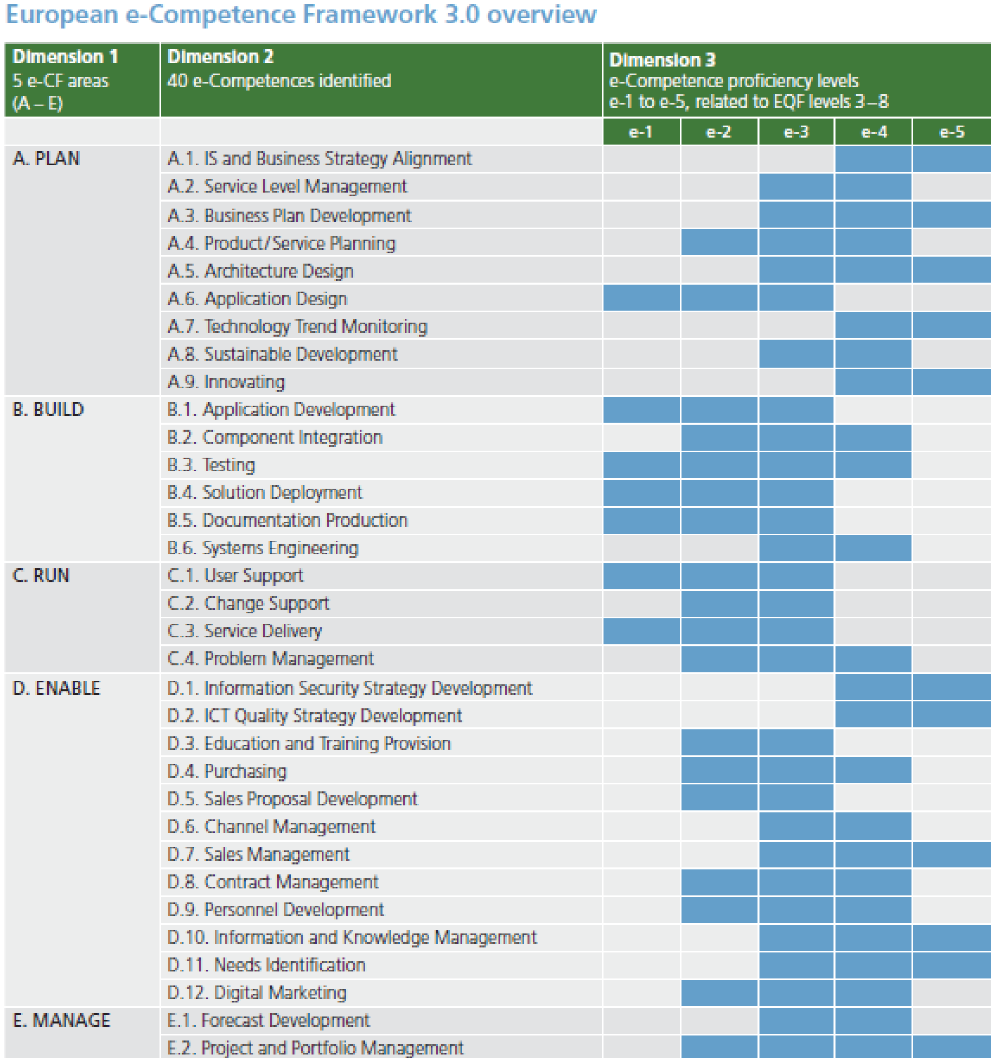
- European e-Competence Framework for IT (e-CFv3.0) dimension
  - Dimension1: 5 competence areas: Plan, Build, Run, Enable, Manage
  - Dimension2: 32 e-competence definition
  - Dimension3: 5 proficiency levels
  - Dimension4: Knowledge and skills


EU e-Competence Framework
Dimension 1
5 e-Competence areas
Dimension 2
32 e-Competence definitions
Dimension 3
e-Competence levels 1 – 5
Dimension 4
knowledge and skills
HR Strategy
Job profiles descriptions
Position description, qualification and experience requirements
Employee
Company Needs
Career paths
Individual competence assessment
Annual individual objectives
Individual development plan
Training plan
Training and certification offering

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

- EDISON CF-DS will be used for defining DS-BoK and MC-DS, linking organizational functions and required knowledge
- Provide basis for individual (self) training and certification

e-CFv3.0 Internal Structure: Refactoring for CF-DS



European e-Competence Framework 3.0 overview

Dimension 1 5 e-CF areas (A – E)	Dimension 2 40 e-Competences identified	Dimension 3 e-Competence proficiency levels e-1 to e-5, related to EQF levels 3–8				
		e-1	e-2	e-3	e-4	e-5
A. PLAN	A.1. IS and Business Strategy Alignment				■	■
	A.2. Service Level Management			■	■	
	A.3. Business Plan Development			■	■	■
	A.4. Product/Service Planning		■	■	■	
	A.5. Architecture Design			■	■	■
	A.6. Application Design	■	■	■		
	A.7. Technology Trend Monitoring				■	■
	A.8. Sustainable Development			■	■	
	A.9. Innovating				■	■
B. BUILD	B.1. Application Development	■	■	■		
	B.2. Component Integration		■	■	■	
	B.3. Testing	■	■	■	■	
	B.4. Solution Deployment	■	■	■		
	B.5. Documentation Production	■	■	■		
	B.6. Systems Engineering			■	■	
C. RUN	C.1. User Support	■	■	■		
	C.2. Change Support		■	■		
	C.3. Service Delivery	■	■	■		
	C.4. Problem Management		■	■	■	
D. ENABLE	D.1. Information Security Strategy Development				■	■
	D.2. ICT Quality Strategy Development				■	■
	D.3. Education and Training Provision		■	■		
	D.4. Purchasing		■	■	■	
	D.5. Sales Proposal Development		■	■		
	D.6. Channel Management			■	■	
	D.7. Sales Management			■	■	■
	D.8. Contract Management		■	■	■	
	D.9. Personnel Development		■	■	■	
	D.10. Information and Knowledge Management			■	■	■
	D.11. Needs Identification			■	■	■
	D.12. Digital Marketing		■	■	■	
E. MANAGE	E.1. Forecast Development			■	■	
	E.2. Project and Portfolio Management		■	■	■	■

- 4 **Dimensions**
  - Competence Areas
  - Competences
  - Proficiency levels
  - Skills and Knowledge
- **5 Competence Areas** defined by ICT Business Process stages
  - Plan
  - Build
  - Deploy
  - Run
  - Manage

-> Refactor to Scientific Research (or Scientific Data) Lifecycle

- See example of RI manager at IG-ETRD wiki and meeting

- Each competence has 5 proficiency levels
  - Ranging from technical to engineering to management to strategist/expert level
- Knowledge and skills property are defined for/by each competence and proficiency level (not unique)

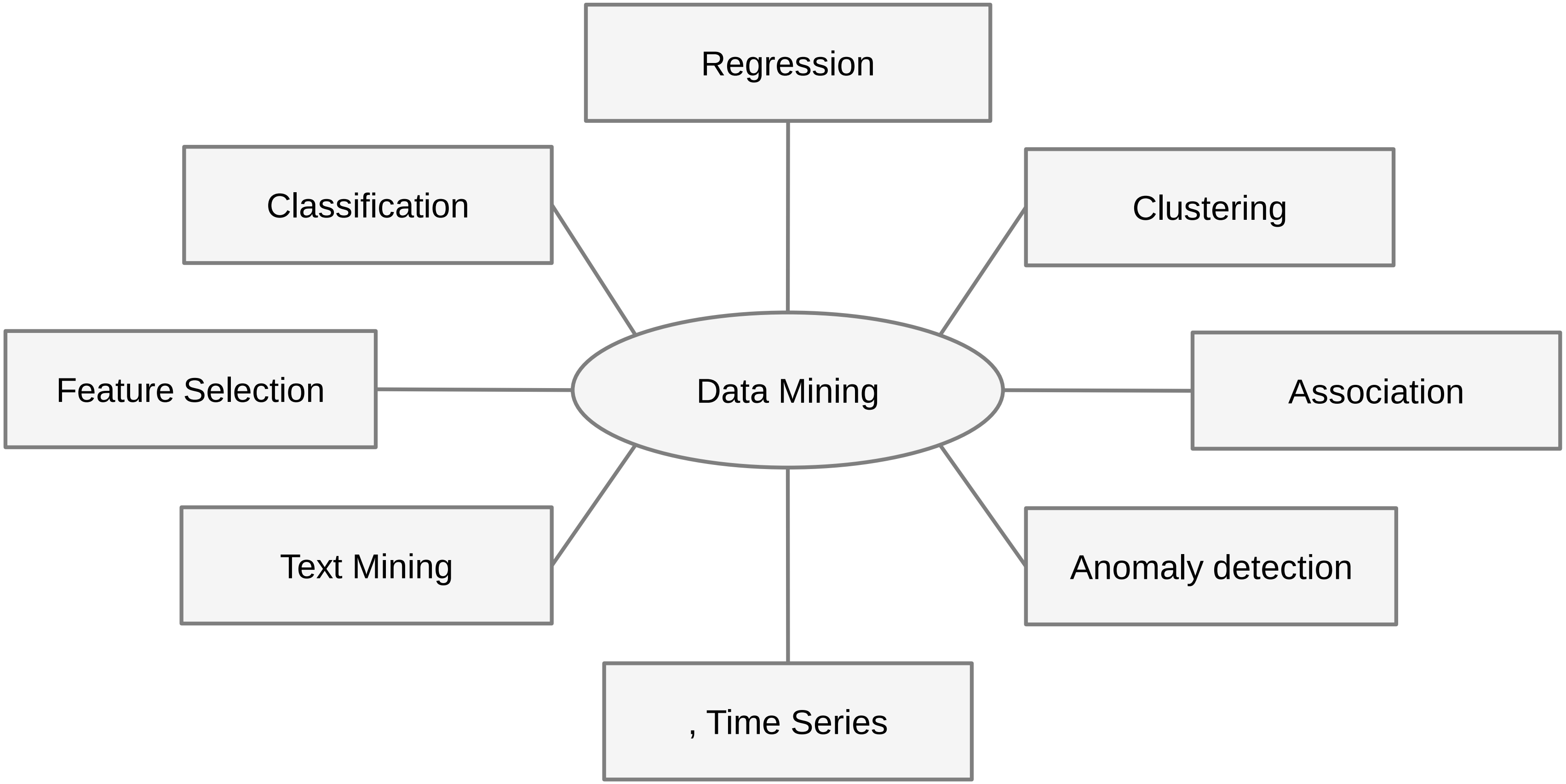
How to become a Data Scientist

- 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

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

Regression

Classification

Clustering

Feature Selection

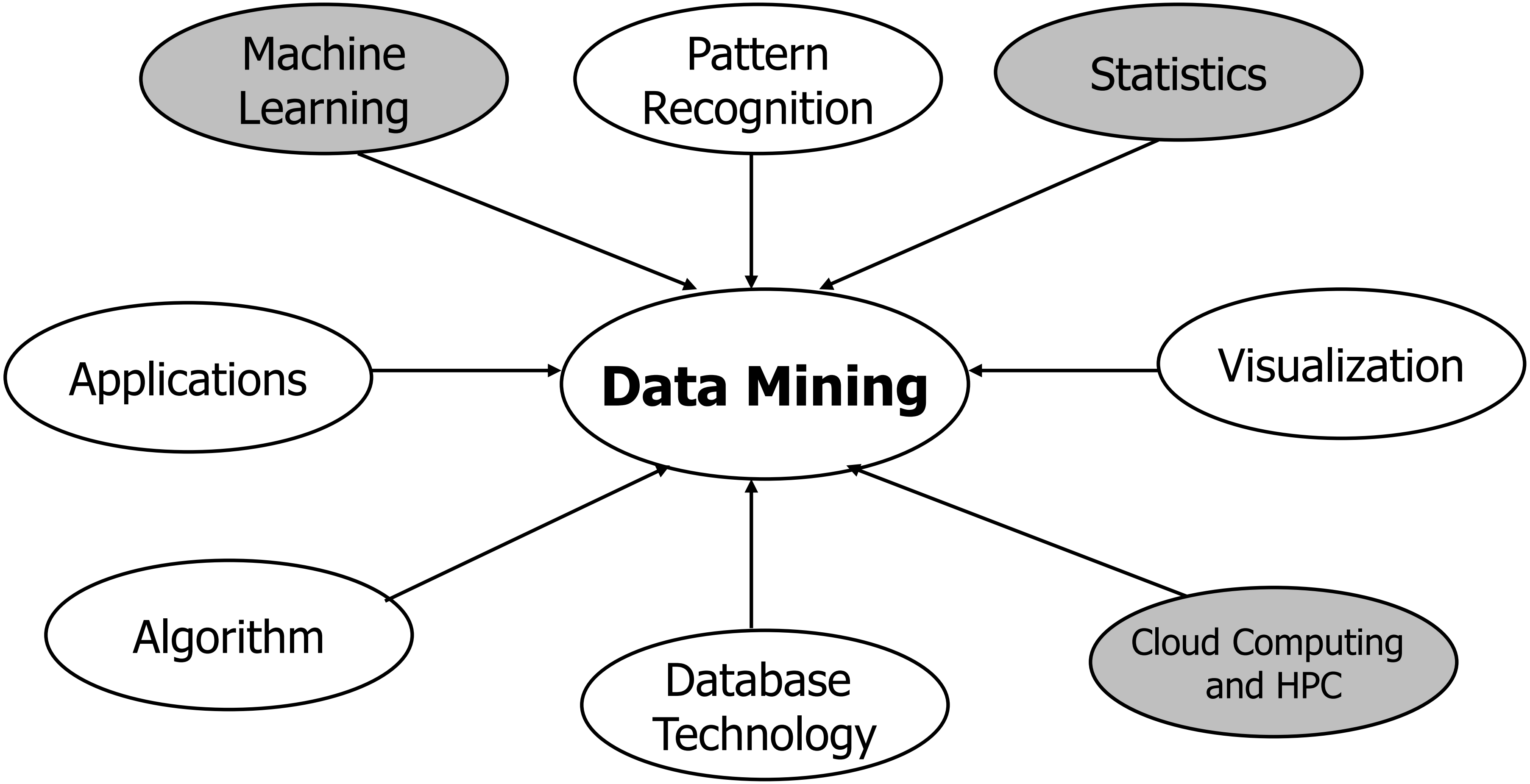
Association

Text Mining

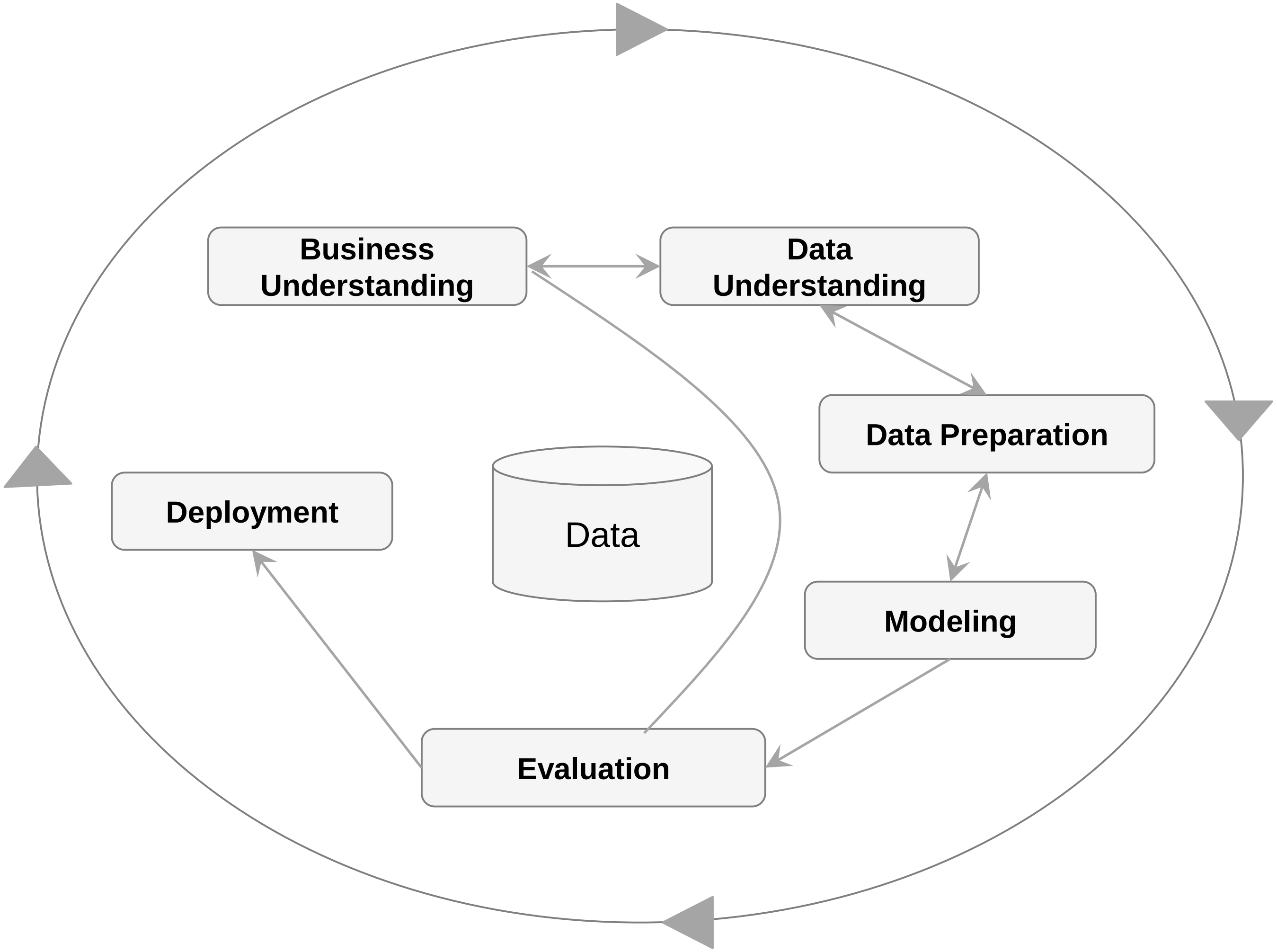
Anomaly detection

, Time Series

Data Mining: Confluence of Multiple Disciplines

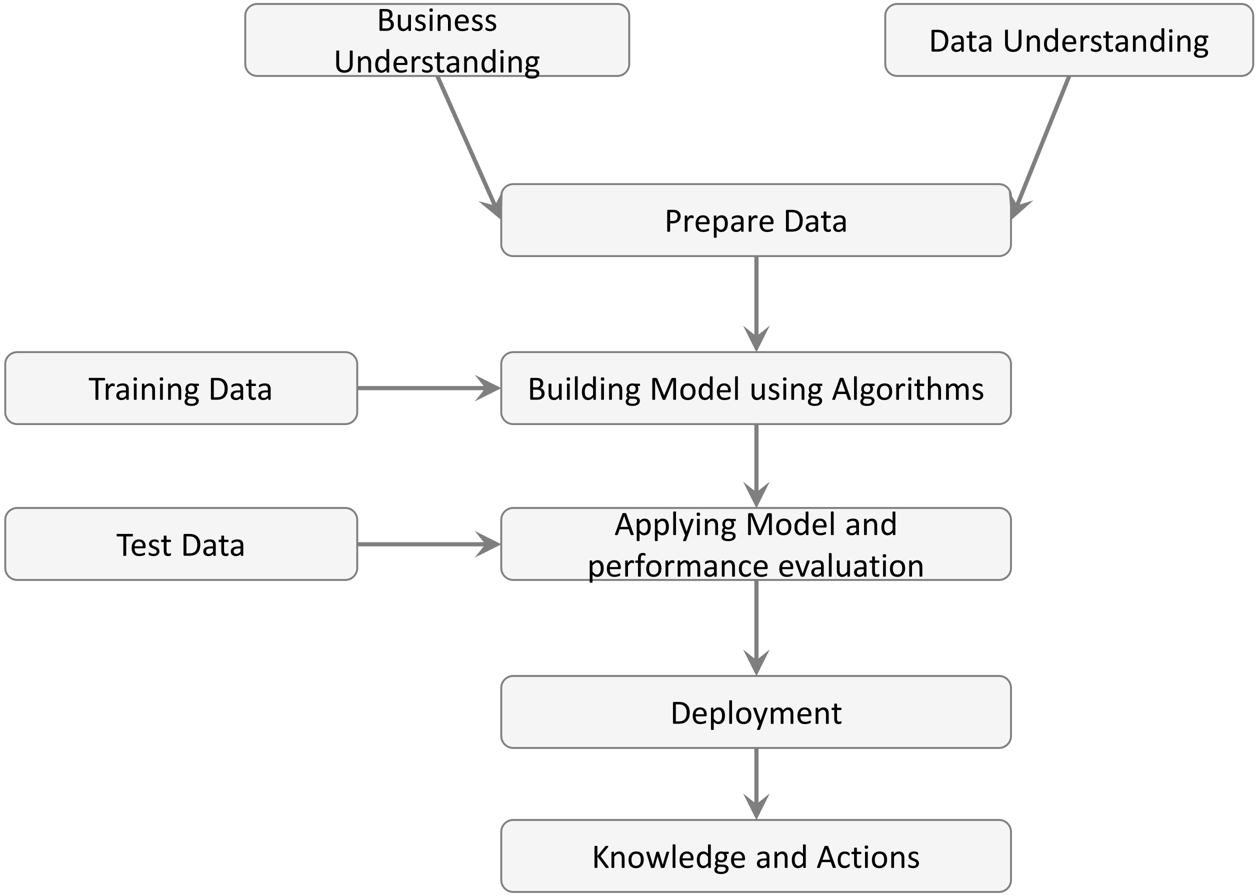


CRISP DM process: Processes and Data Lifecycle



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

Process of Data Analysis



Business Understanding

Data Understanding

Prepare Data

Training Data

Building Model using Algorithms

Test Data

Applying Model and performance evaluation

Deployment

Knowledge and Actions

1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge