Becoming a Data Scientist:
(Teaching Data Scientists)

How develop Data Science and Analytics related competences and professional skills: Foundations and Recommendations

EDISON
building the data science profession

Yuri Demchenko, University of Amsterdam

EDISON Project and Initiative

19 October 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

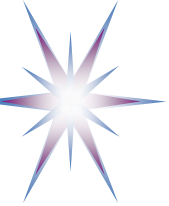
- Background: Data driven research and demand for new skills
  - Foundation, recent reports, studies and facts

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

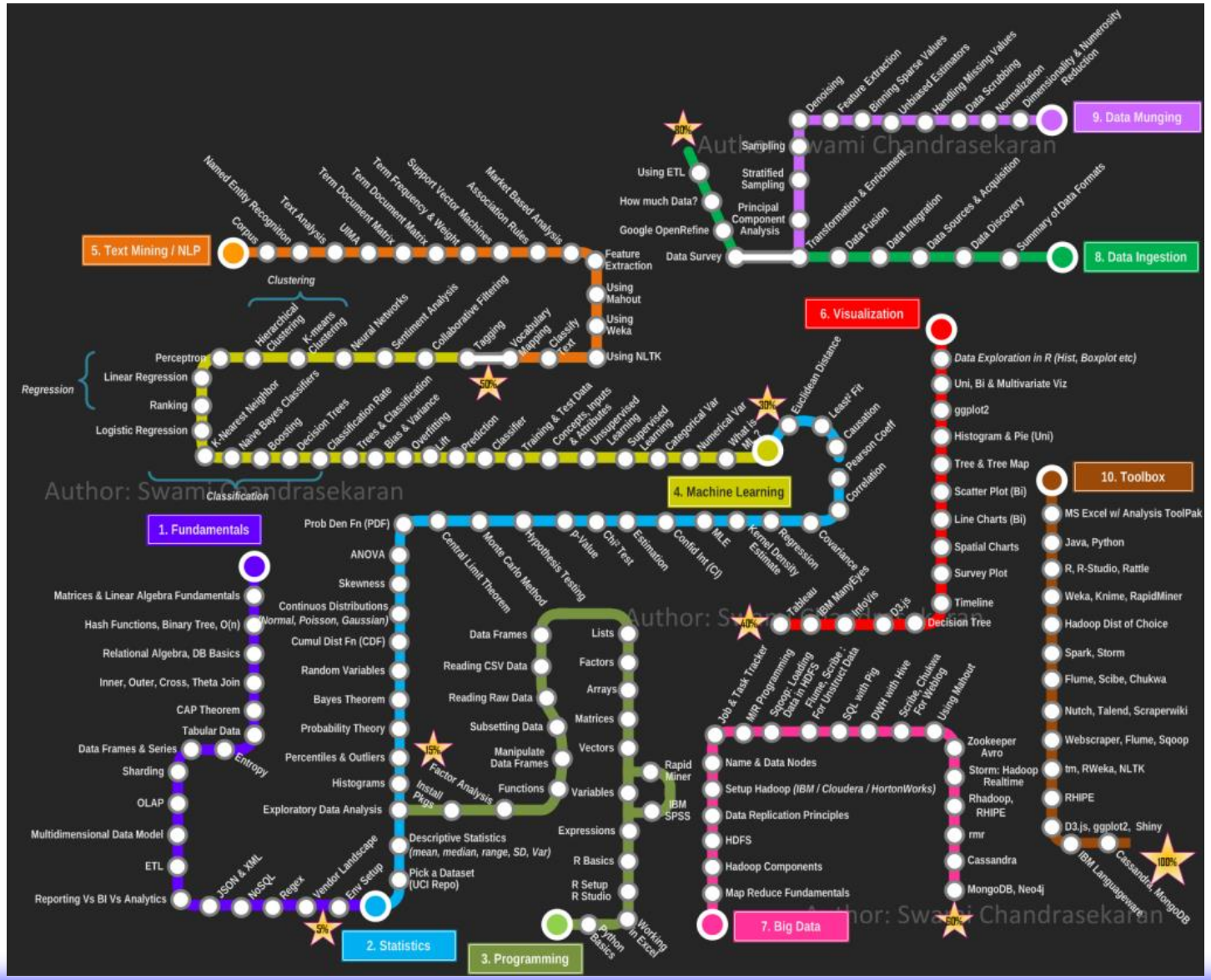
- Use of EDSF and Example curricula
  - Competences assessment
  - Building Data Science team
- **Discussion**



Yuri Demchenko, Senior Researcher, Lecturer, UvA

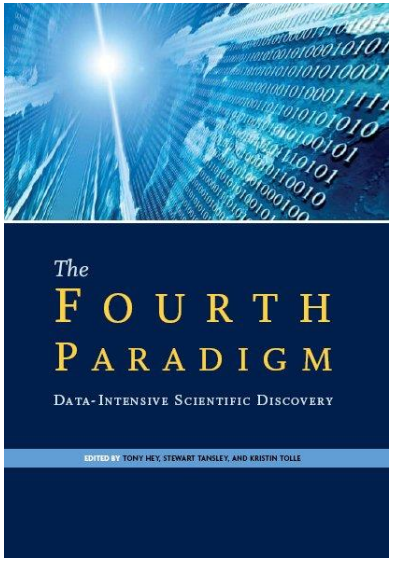
- Graduated and PhD from National Technical University of Ukraine “Kiev Polytechnic Institute”
  - University of Amsterdam – since 2003
- Research areas
  - Big Data Infrastructure and Data Science platforms
  - Cloud architecture, cloud automation and DevOps
  - Cloud security and compliance
- Teaching courses (on campus and online)
  - Big Data Infrastructure and Technologies
  - Cloud powered Software Engineering and DevOps
  - Data Science Foundations, Professional Issues in Data Science
  - Security Engineering
- Recent projects
  - EDISON: Building the Data Science Profession for Europe
  - MATES: Digitalisation of the European Blue Economy
  - CYCLONE: Multi-cloud automation platform for cloud based applications
  - GEANT4 Research: Cloud aware networking infrastructure provisioning on-demand

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

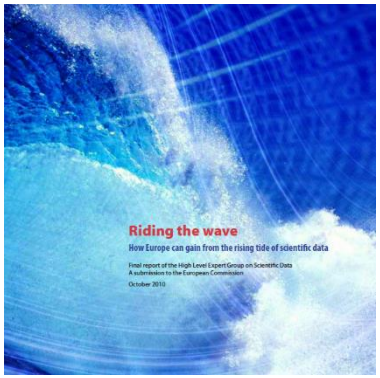
Visionaries and Drivers: Seminal works, High level reports, Activities



The Fourth Paradigm: Data-Intensive Scientific Discovery.

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

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



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

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

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

RDA RESEARCH DATA ALLIANCE

Research Data Sharing without barriers

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

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

An RDA Europe Report. December 2014

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

HLEG report on European Open Science Cloud

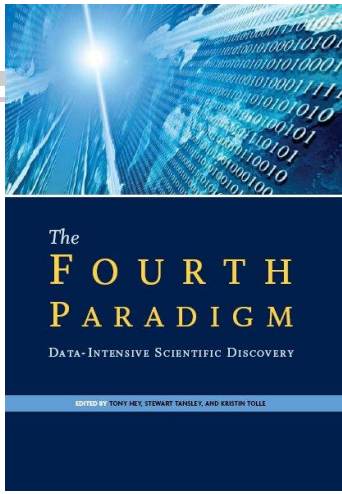
(October 2016)

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

Emergence of Cognitive Technologies

(IBM Watson, Cortana and others)

The Fourth Paradigm of Scientific Research



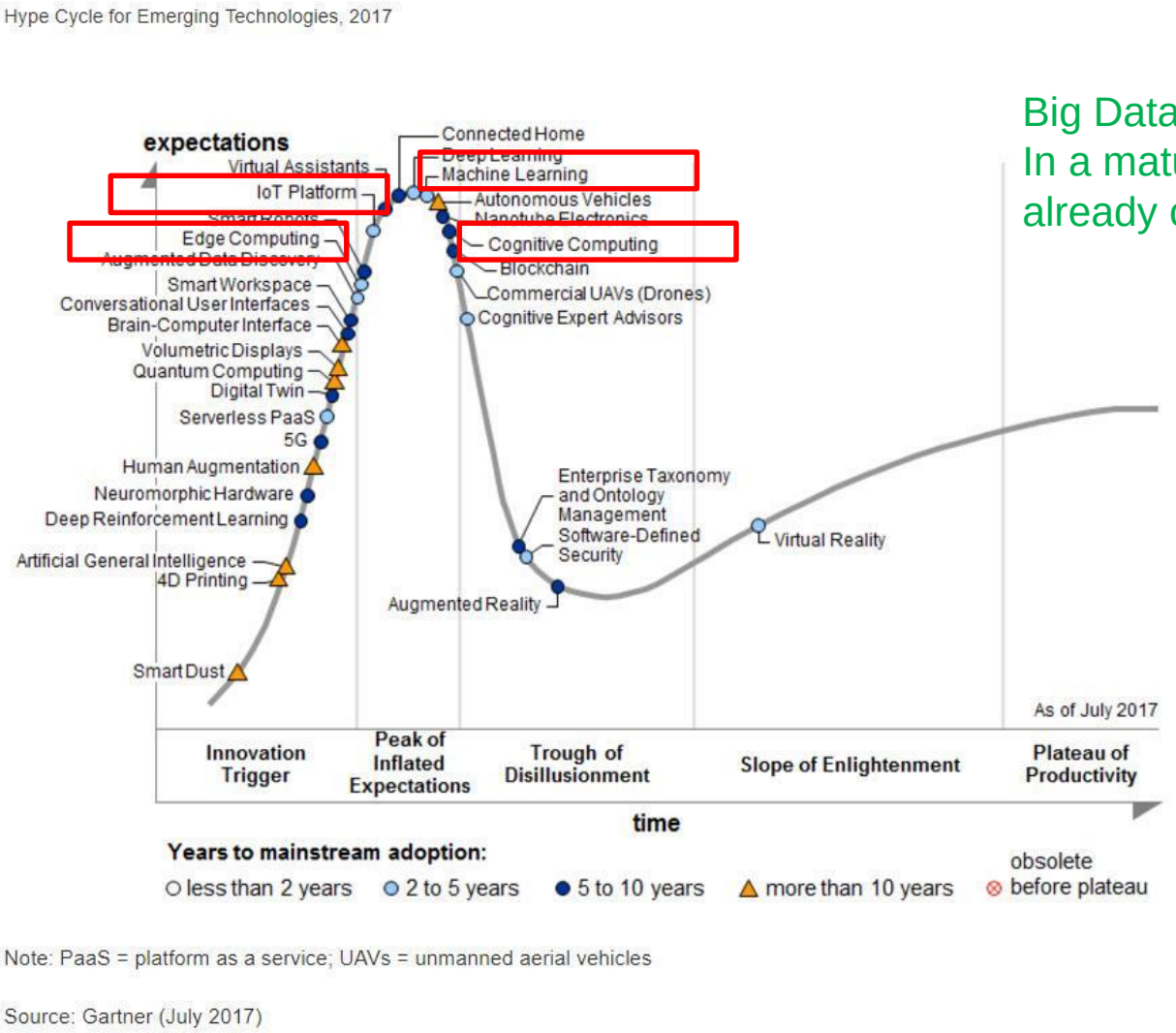
1. Theory, hypothesis and logical reasoning
2. Observation or Experiment, e.g.
   - Newton observed apples falling to design his theory of mechanics
   - Gallileo Galilei made experiments with falling objects from the Pisa leaning tower
3. Simulation of theory or model
   - Digital simulation can prove 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

HLEG EOSC Report Essentials – **Core Data Experts** [ref]

- **Core Data Experts** is a new class of colleagues with core scientific professional competencies and the communication skills to fill the gap between the two cultures.
  - **Core data experts** are neither computer savvy research scientists nor are they hard-core data or computer scientists or software engineers.
  - They should be technical data experts, though proficient enough in the content domain where they work routinely from the very beginning (experimental design, proposal writing) until the very end of the data discovery cycle
  - Converge two communities:
    - Scientists need to be educated to the point where they hire, support and respect Core Data Experts
    - Data Scientists (Core Data Experts) need to bring the value to scientific research and organisations
- Implementation of the EOSC needs to include instruments to help train, retain and recognise this expertise,
  - In order to support the 1.7 million scientists and over 70 million people working in innovation.

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

Gartner Technology Hypercycle (August 2017)



Big Data and Cloud Computing:
In a maturity stage –
already commodity services

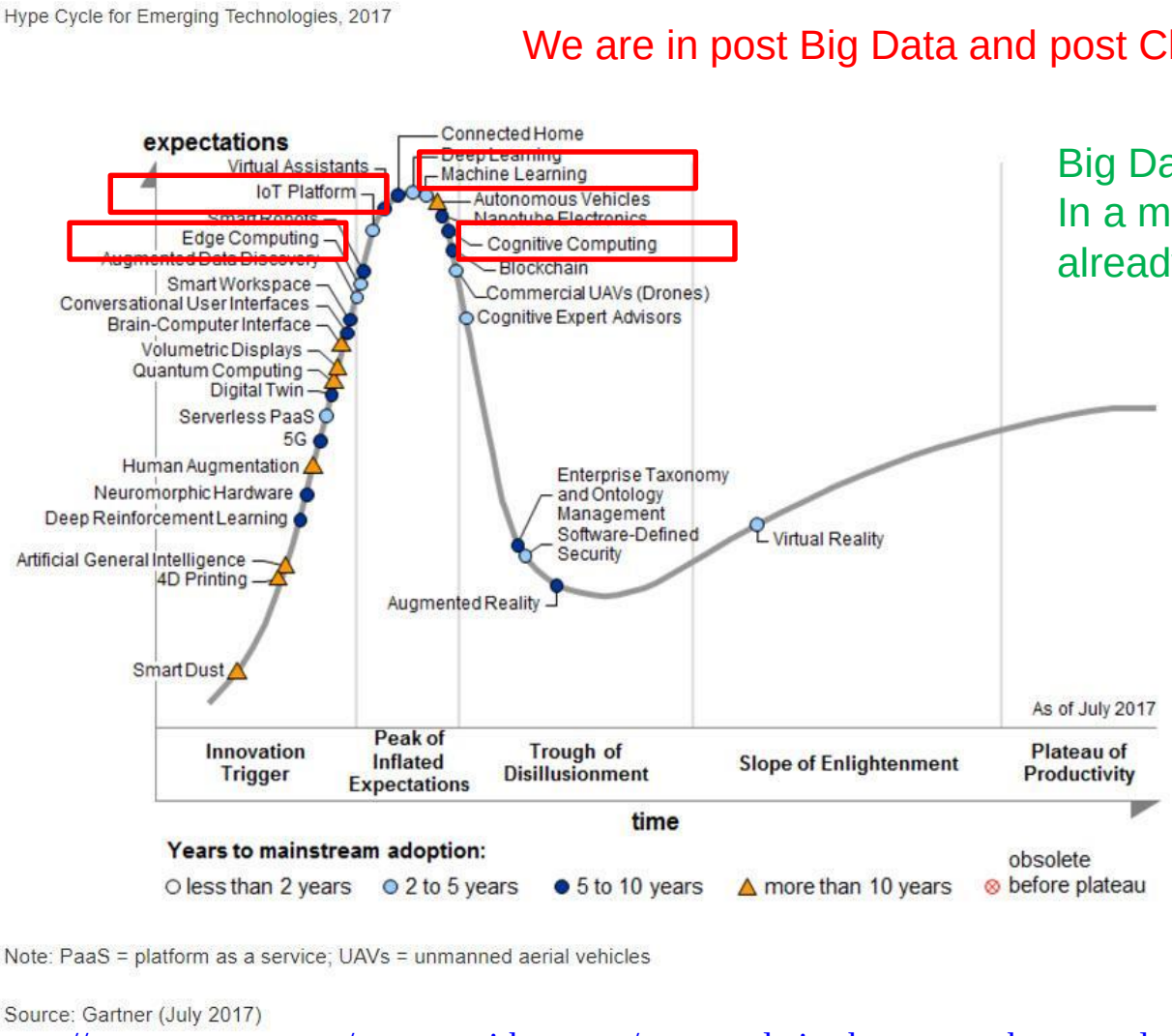
Note: PaaS = platform as a service; UAVs = unmanned aerial vehicles

Source: Gartner (July 2017)

[ref] https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/

Gartner Technology Hypercycle (August 2017)

We are in post Big Data and post Cloud Computing stage



Big Data and Cloud Computing:
In a maturity stage –
already commodity services

Note: PaaS = platform as a service; UAVs = unmanned aerial vehicles

Source: Gartner (July 2017)

[ref] https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/

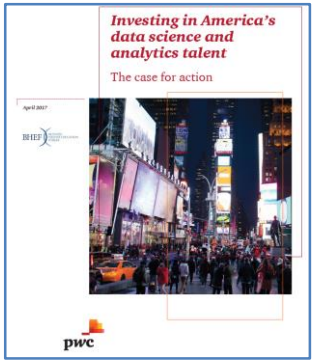
International and EU studies on data-driven skills

Industry reports on Data Science Analytics and Data enabled skills demand

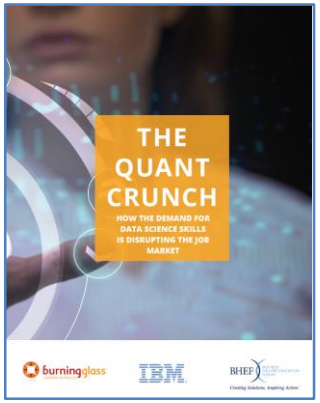
- Final Report on European Data Market Study by IDC (Feb 2017)
  - The EU data market in 2016 estimated EUR 60 Bln (growth 9.5% from EUR 54.3 Bln in 2015)
    - Estimated EUR 106 Bln in 2020
  - Number of data workers 6.1 mln (2016) - increase 2.6% from 2015
    - Estimated EUR 10.4 million in 2020
  - 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

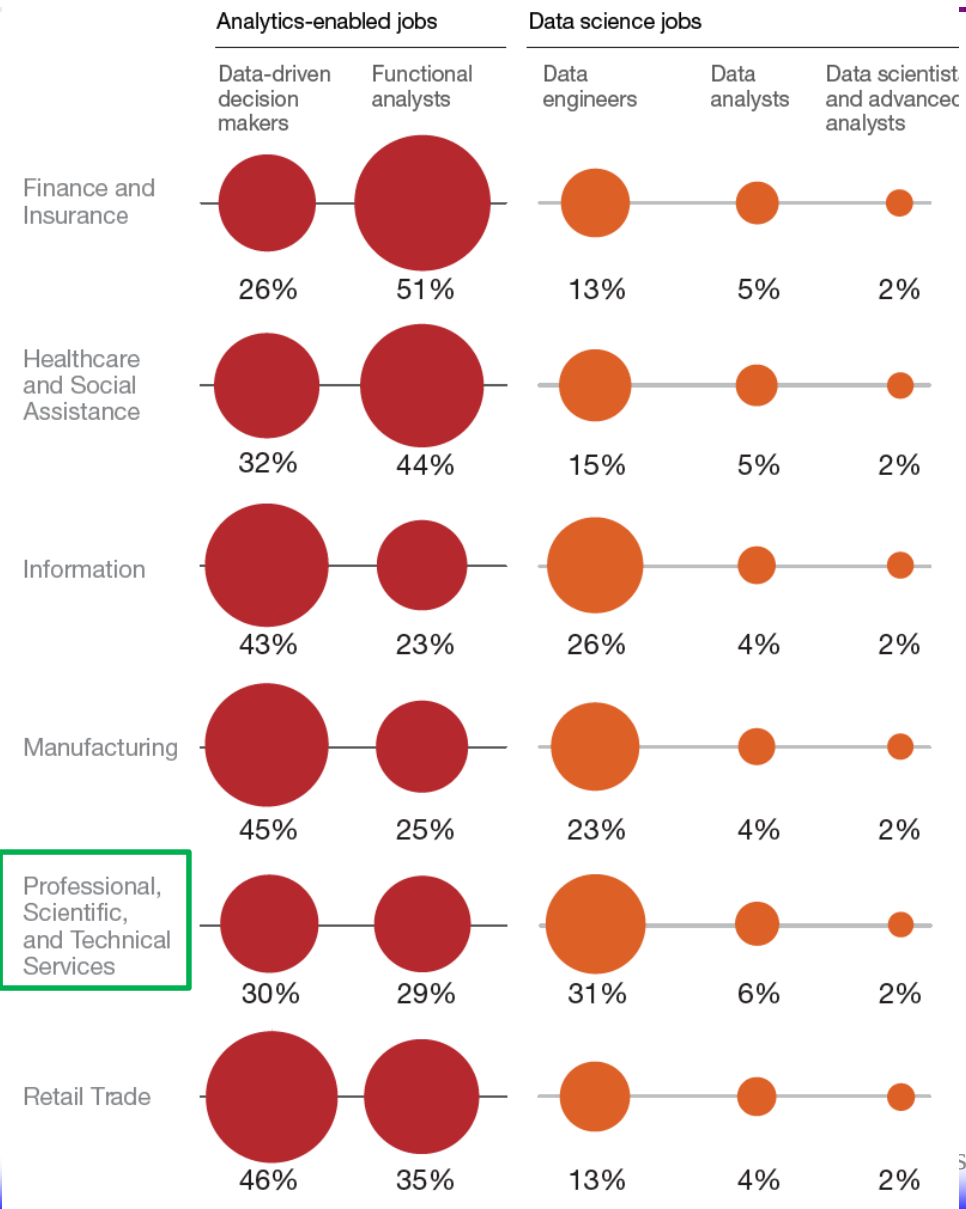


Citing EDISON and EDSF



Influenced by EDISON

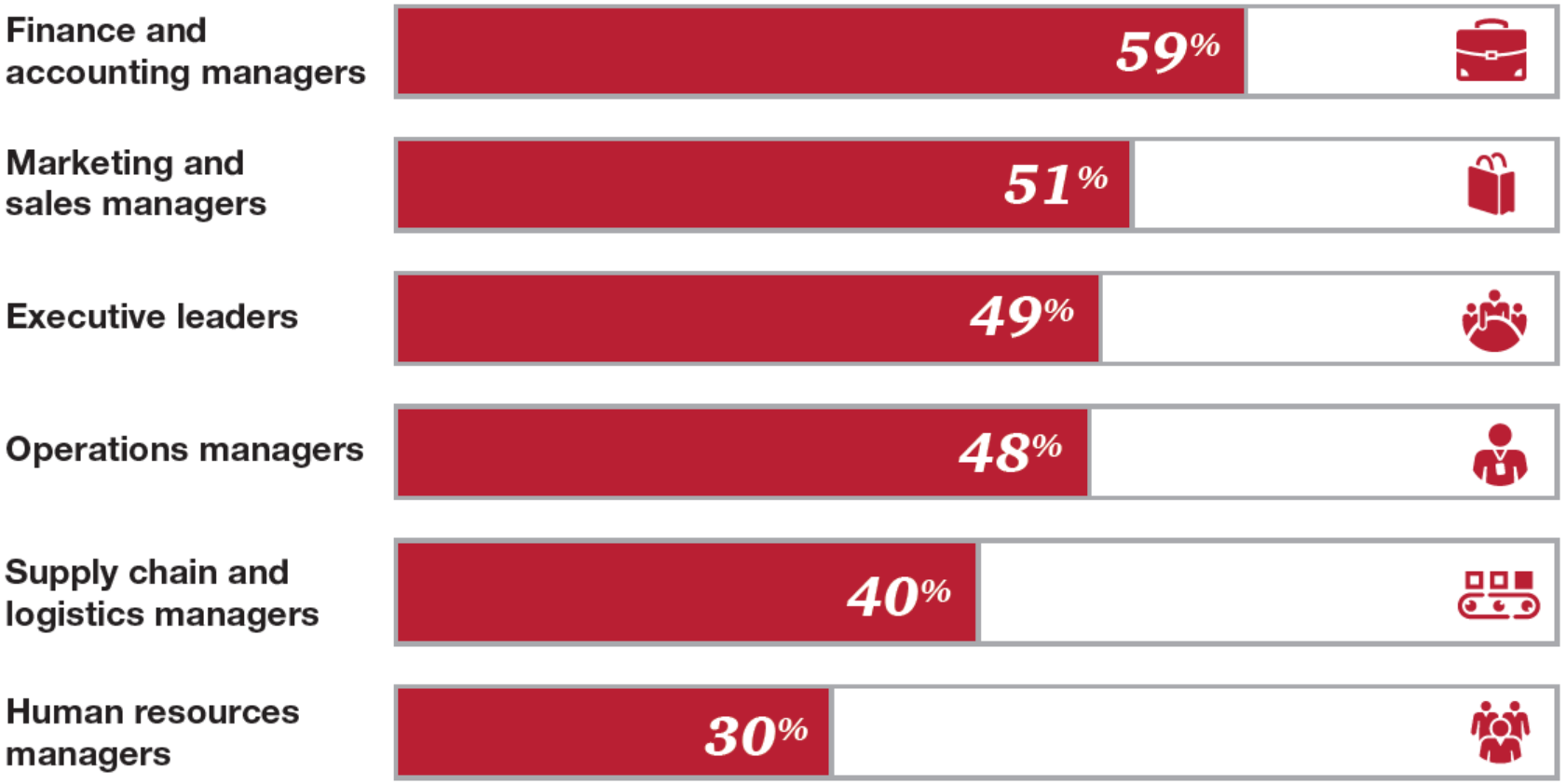
PwC&BHEF: Demand for DSA enabled jobs



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

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

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

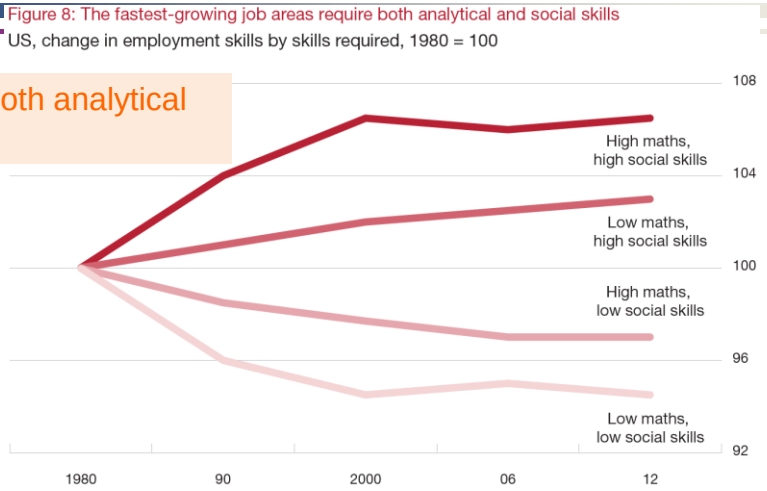


Percent of employers who say data science and analytics skills will be 'required of all managers' by 2020

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

PwC&BHEF: Skills that are tough to find

Faster growing jobs require both analytical and social skills



97% Cybersecurity

95% Data science and analytics

83% Critical thinking and problem solving

79% Design/ systems thinking

79% Innovation and creativity

78% Global perspective

78% Cognitive flexibility

74% Cross-disciplinary ability

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

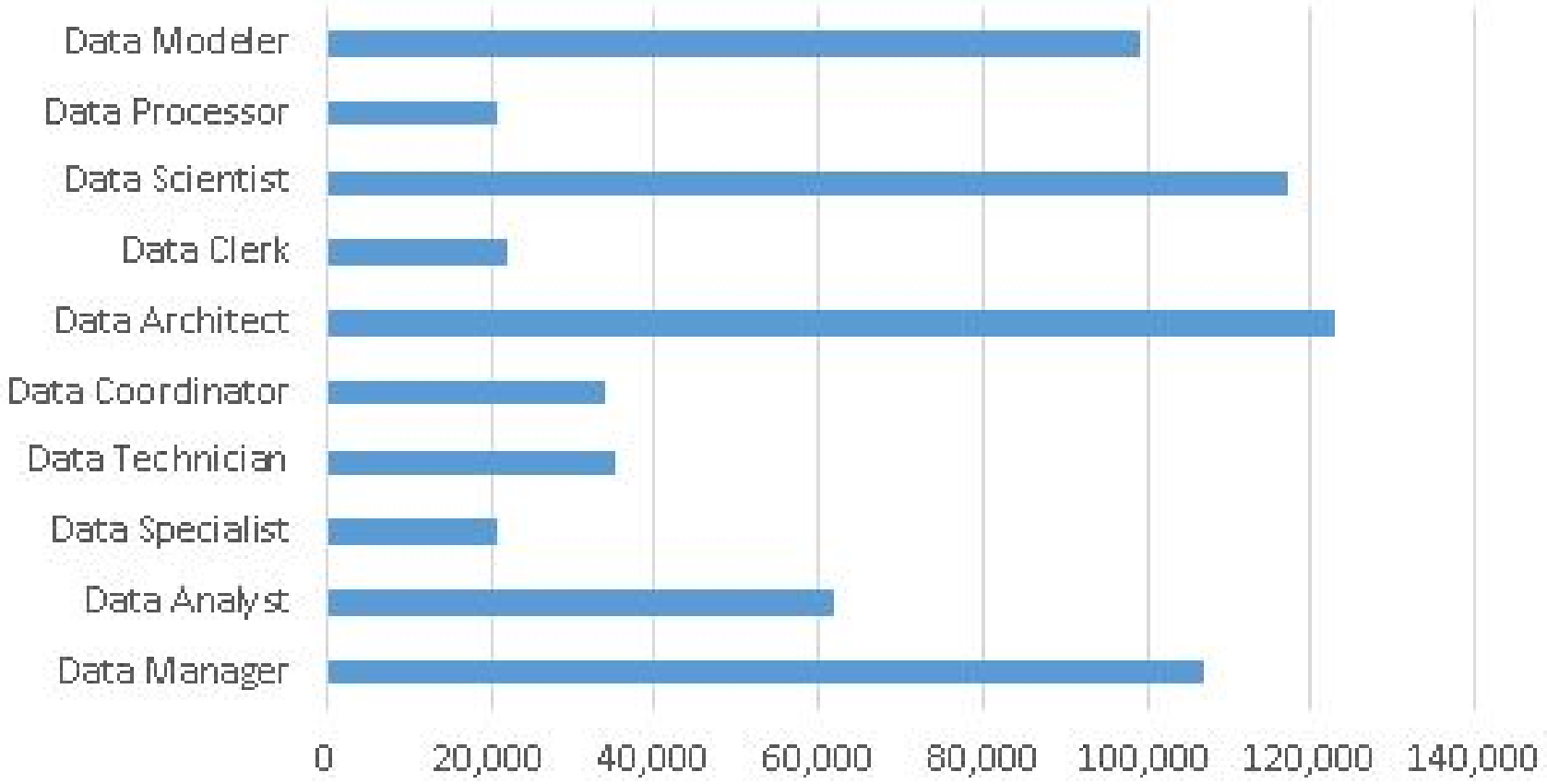
Source: Business Roundtable (2017).

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

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

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

Closer look at Data related Jobs and Salaries (2016)



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

OECD and UN on Digital Economy and Data Literacy

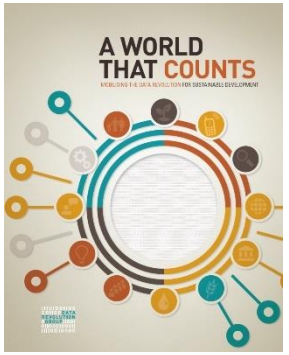
OECD (Organisation for Economic Coopration and Development)

- 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

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

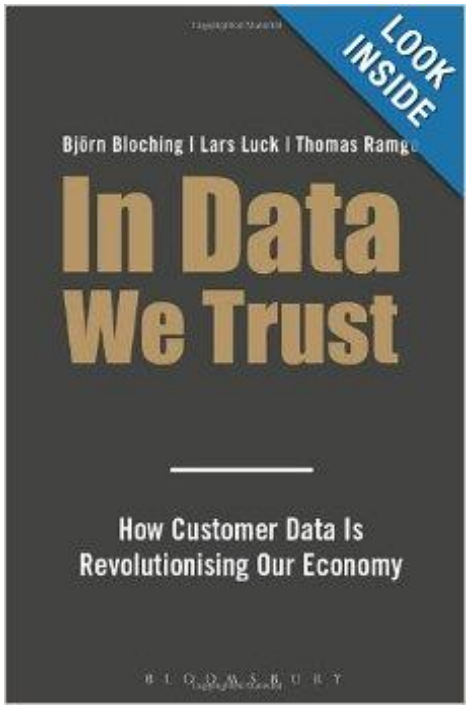
Data Driven Victories and Failures - Politics

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



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

- 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, other facts and Data Analytics

- On-going scandal with Cambridge Analytica

This area intentionally left blank

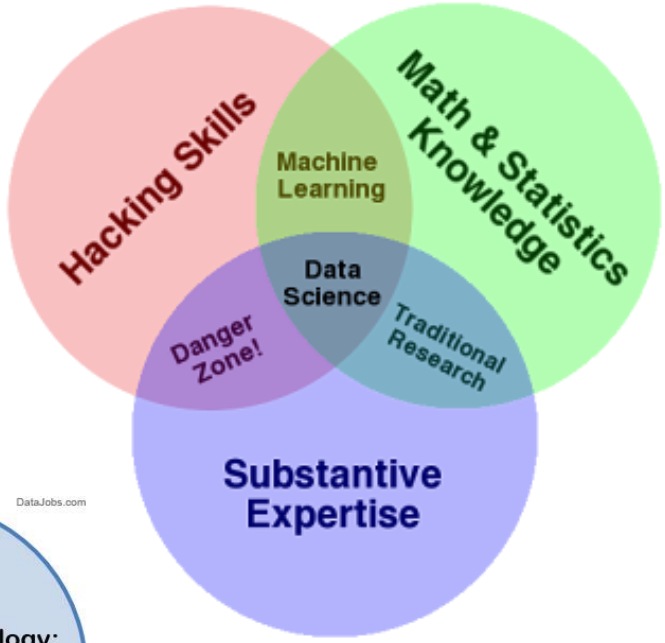
- Growing importance of ethical factor
  - Education is essential to tame new element/dimension of our life – Data
  - Existing ICT Manifesto for ICT Professionals

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

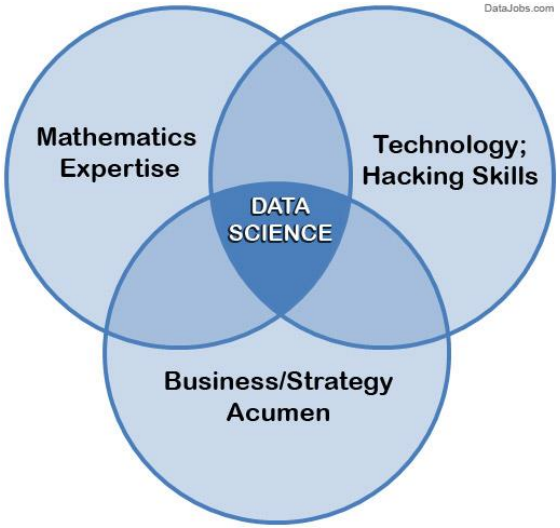
Challenge for Education: Sustainable ICT and Data Skills Development

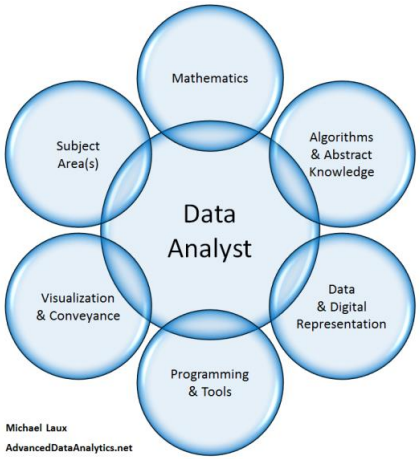
- Educate vs Train
  - Training is a short term solution
  - Education is a basis for sustainable skills development
  - *Importance of workplace or professional attitude skills (not covered in academic curricula)*
- Technology focus changes every 3-4 years
  - Study: 50% of academic curricula are outdated at the time of graduation
- *Growing influence of Big5 technology companies: Amazon, Microsoft, Google, Facebook, Apple*
- 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

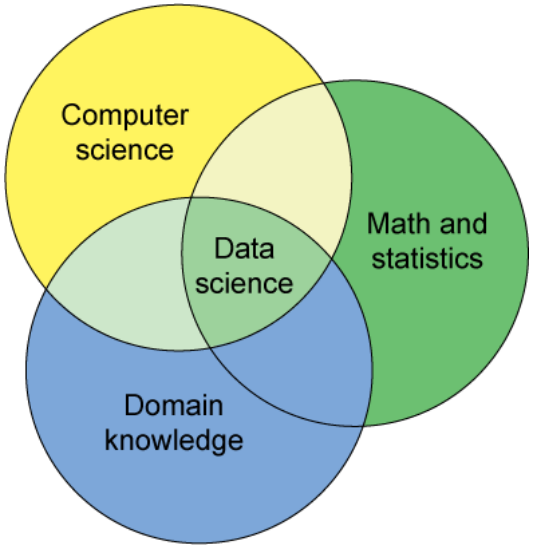
Data Scientist definitions: From Math to Hacking



- Strongly depend on the background of the Data Scientist

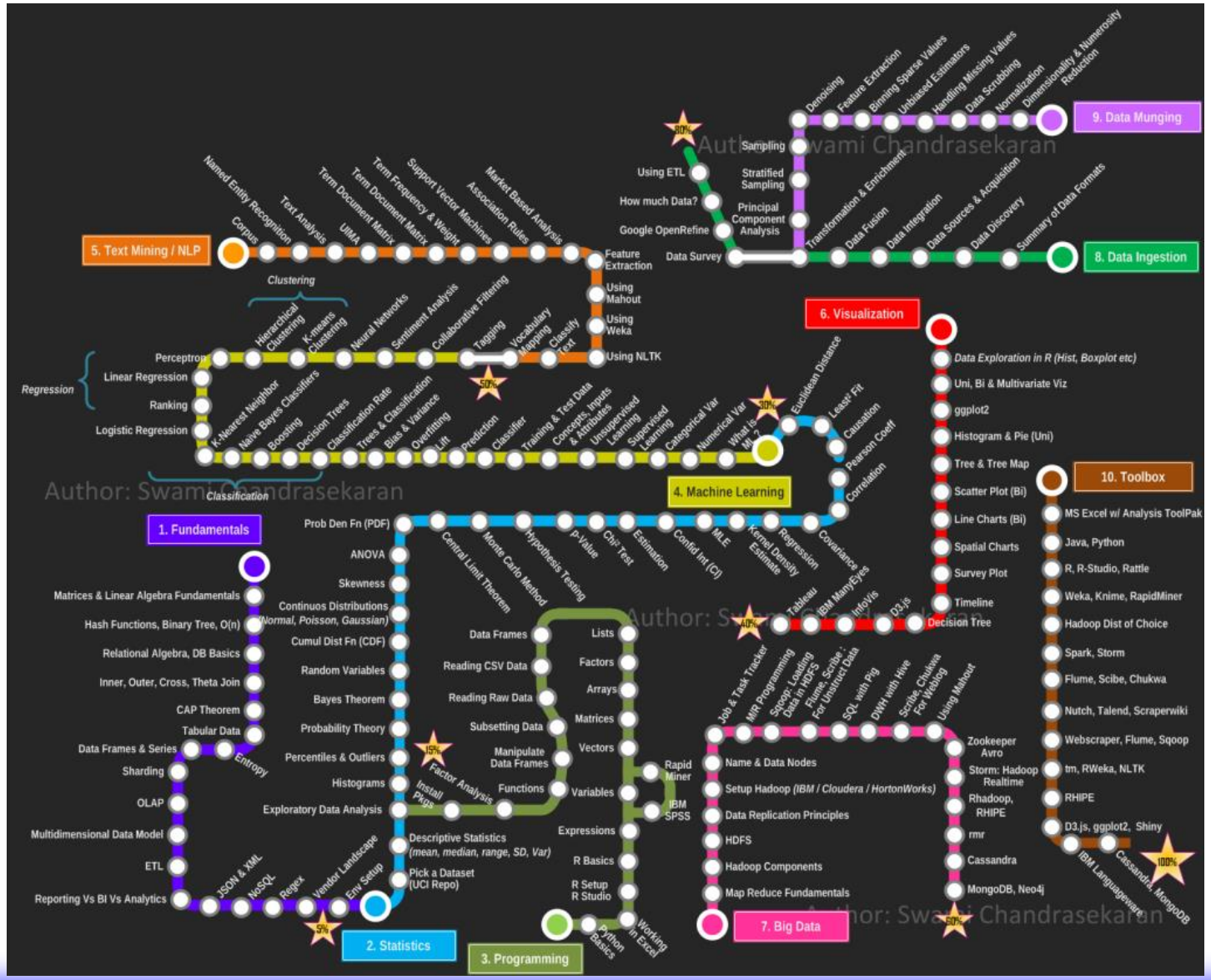






Becoming a Data Scientist by Swami Chandrasekaran (2013)

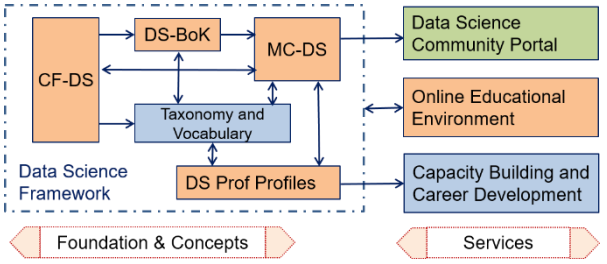
http://nirvacana.com/thoughts/becoming-a-data-scientist/

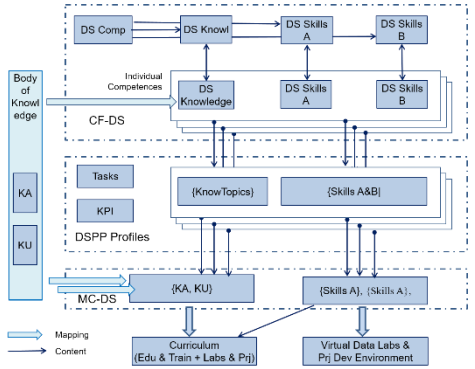


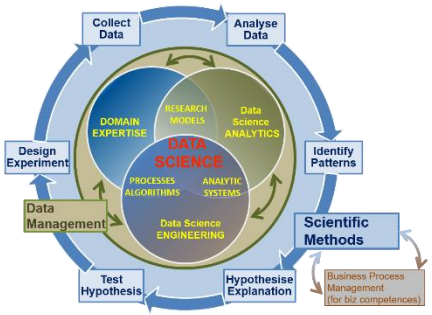
- Good and practical advice how to learn Data Science, step by step
- Follow the route

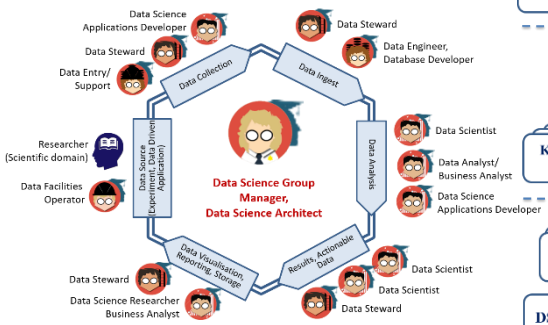
EDISON Products for Data Science Skills Management and Curriculum Design

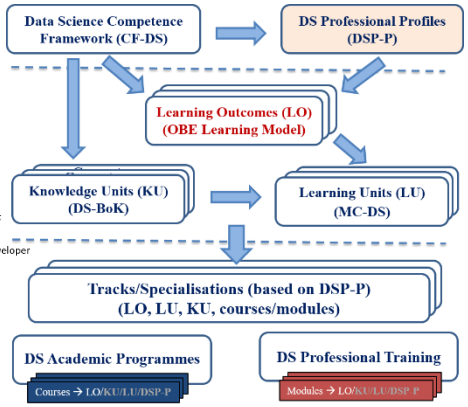
- EDISON Data Science Framework (EDSF)
  - Compliant with EU standards on competences and professional occupations e-CFv3.0, ESCO
  - Customisable courses design for targeted education and training
- Skills development and career management for Core Data Experts and related data handling professions
- 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)



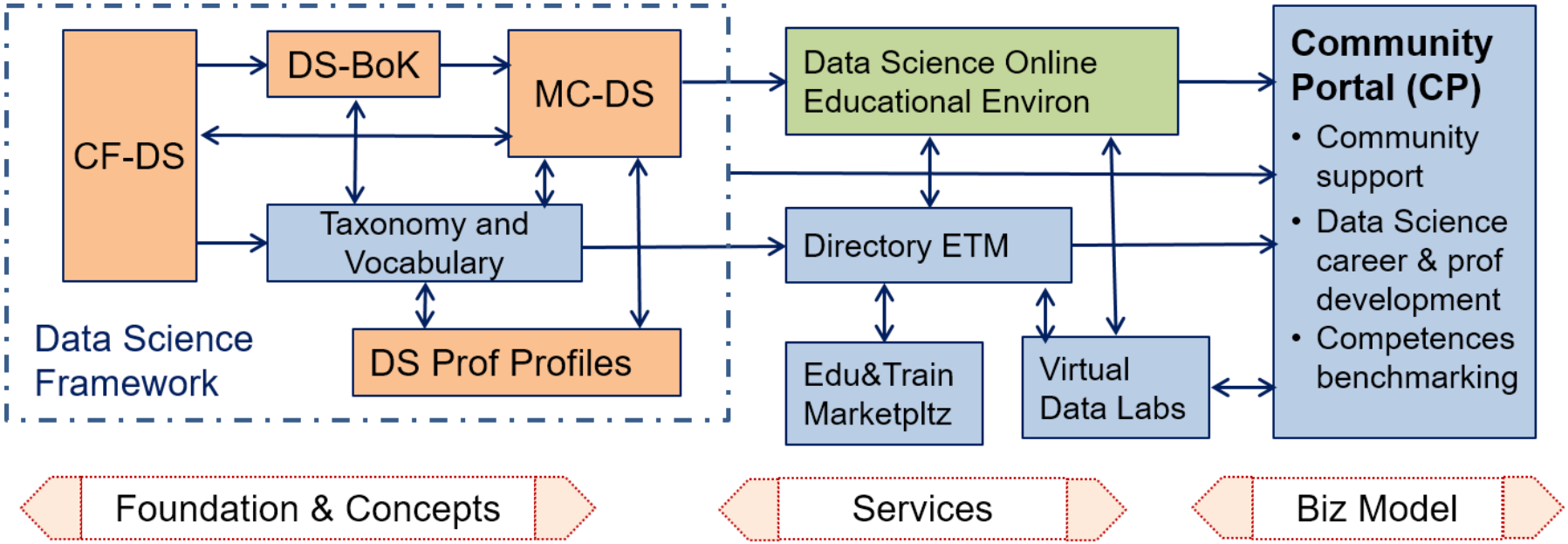








EDISON Data Science Framework (EDSF)



EDISON Framework components

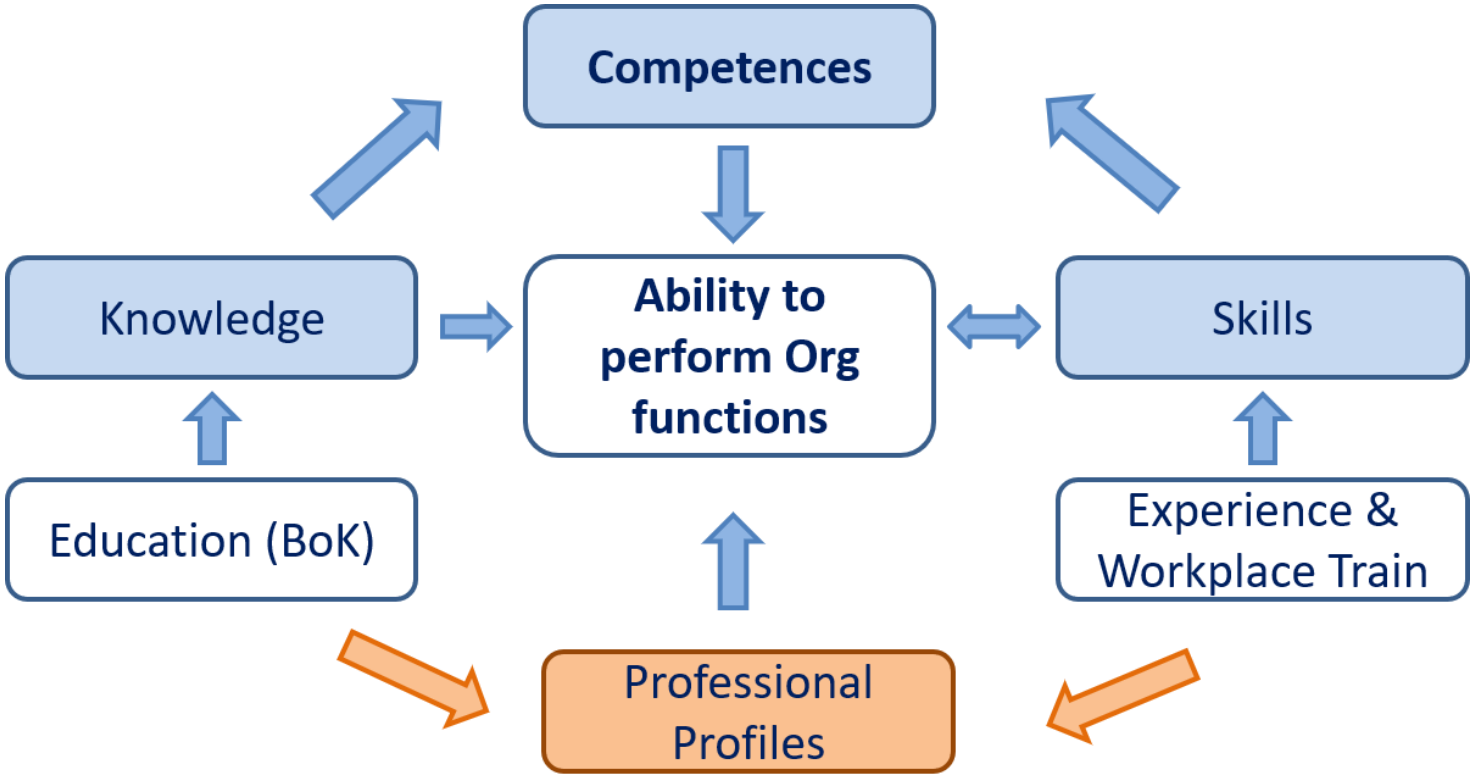
- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- DSP – Data Science Professional profiles
- Data Science Taxonomies and Scientific Disciplines Classification
- EOEE - EDISON Online Education Environment

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.

Competences Map to Knowledge and Skills

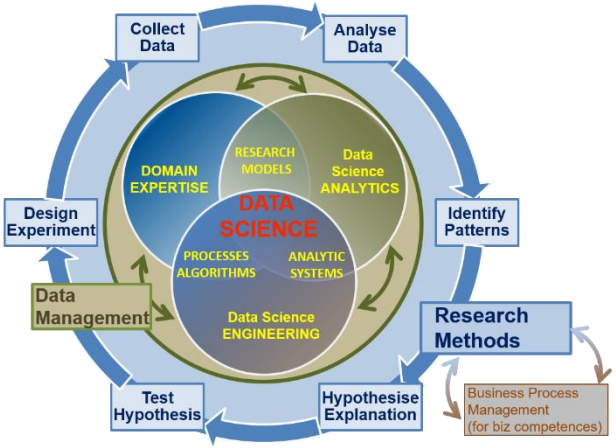
- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results



Data Scientist definition

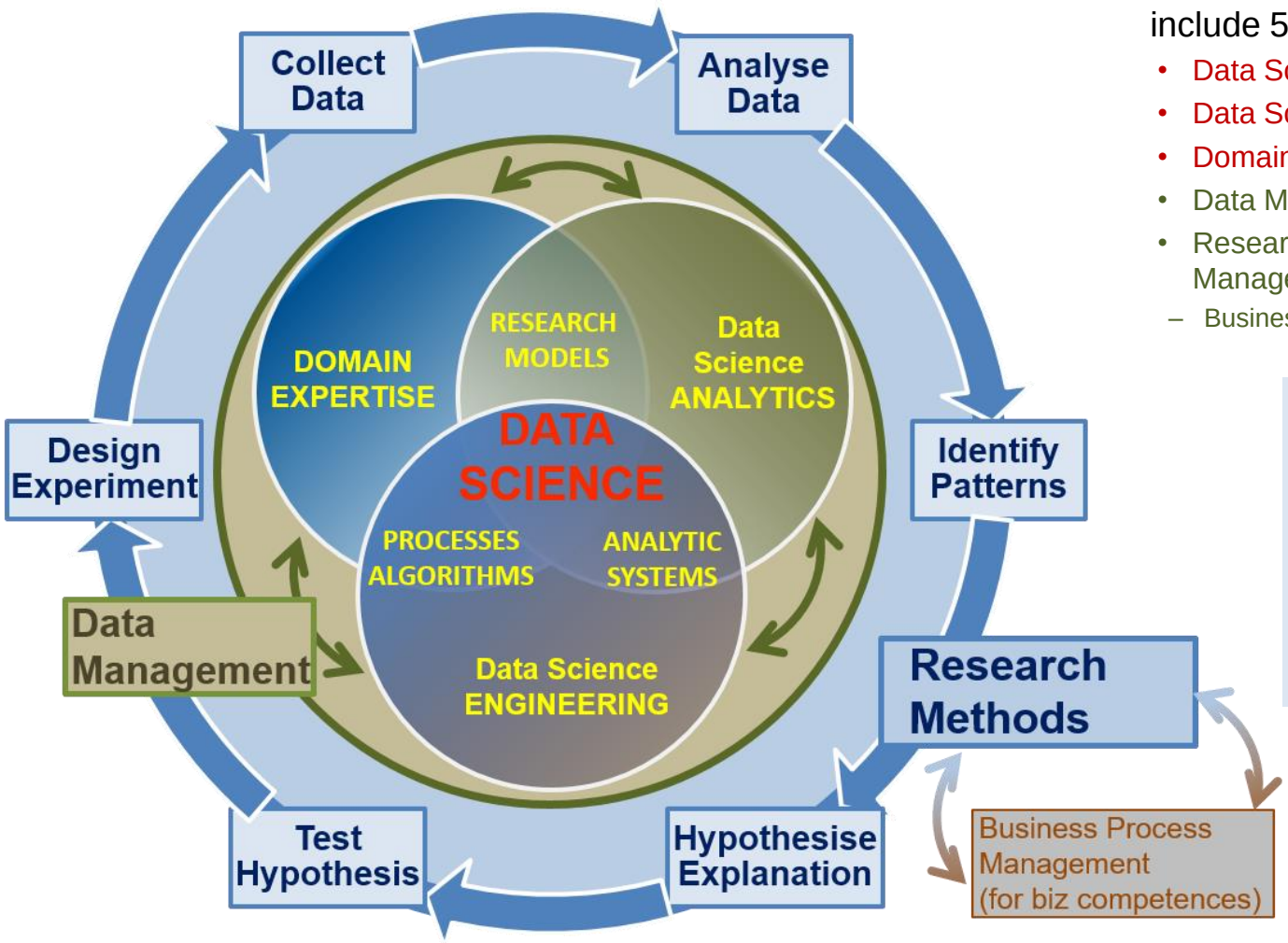
Based on the definitions by NIST SP1500 – 2015, extended by EDISON

- *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)

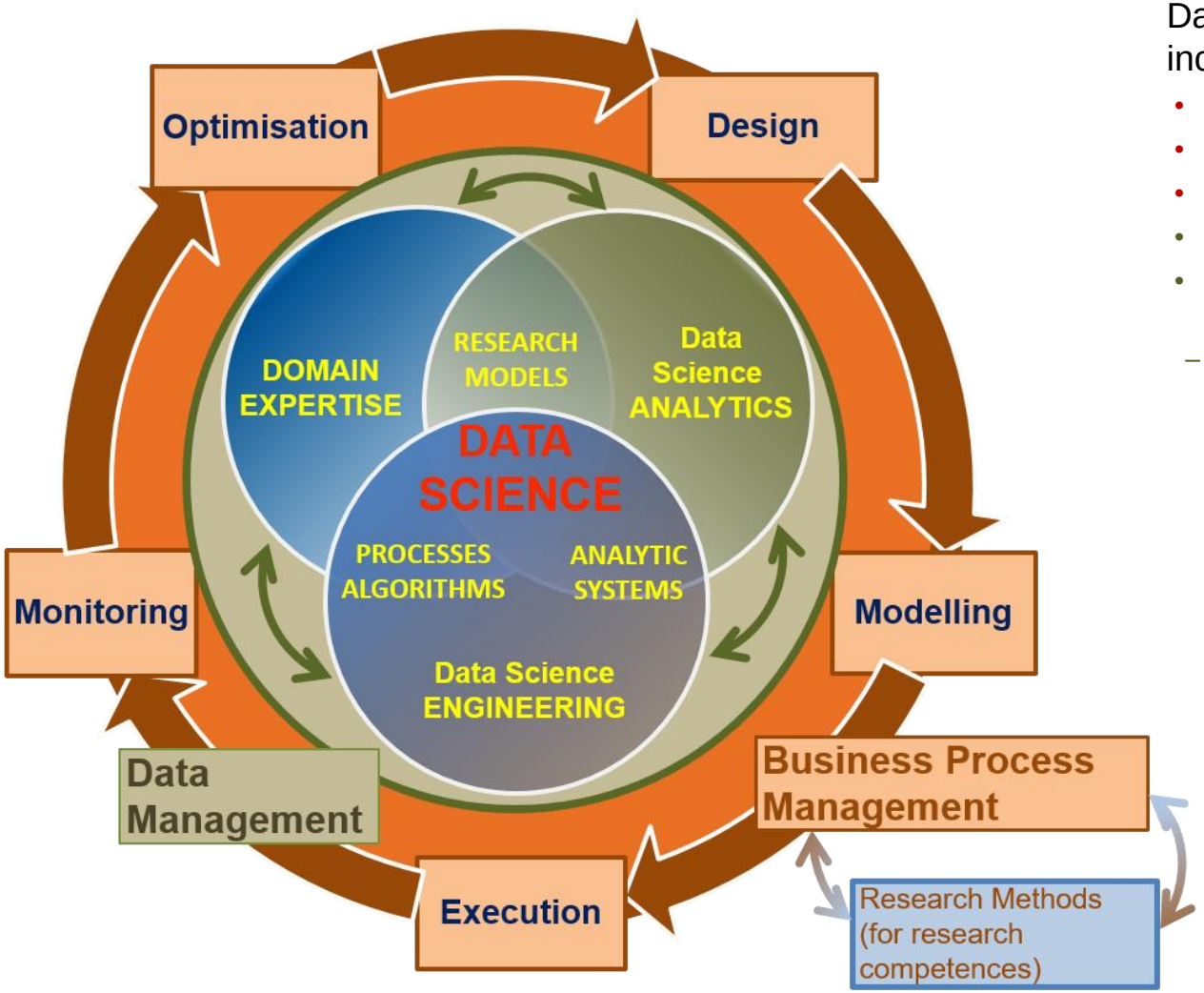
Scientific Methods

- 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

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)

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, initiative, 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

Maritime Industry Digital Transformation and Skills Strategy – Toward Industry 4.0

OCCUPATIONAL PROFILES

CURRICULUM DESIGN

EDUCATION & TRAINING

KNOWLEDGE TRANSFER

LONG-TERM SKILLS STRATEGY

OCEAN LITERACY

Digital Transformation

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

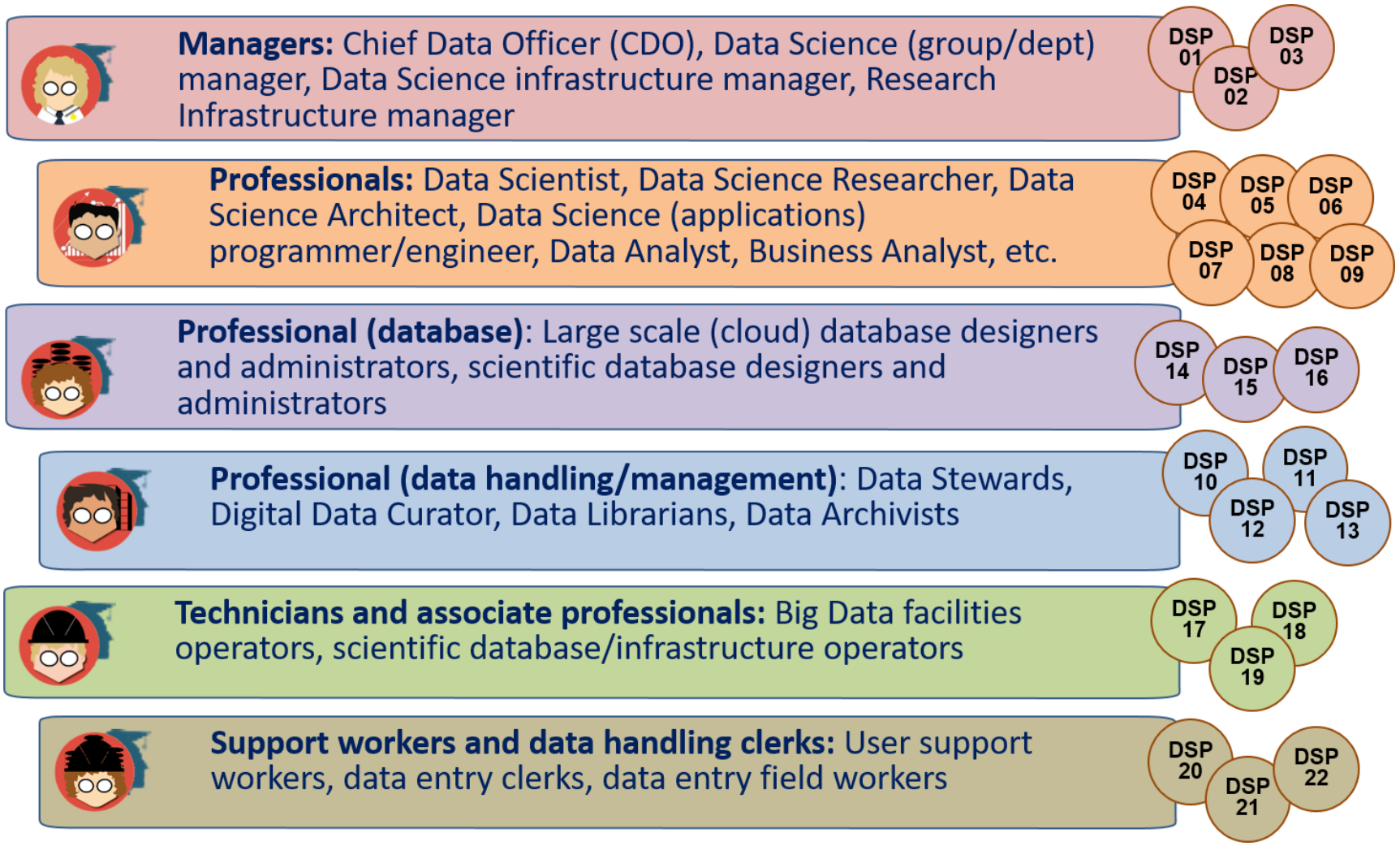
Digital Competences/Skills

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

Practical Application of the CF-DS

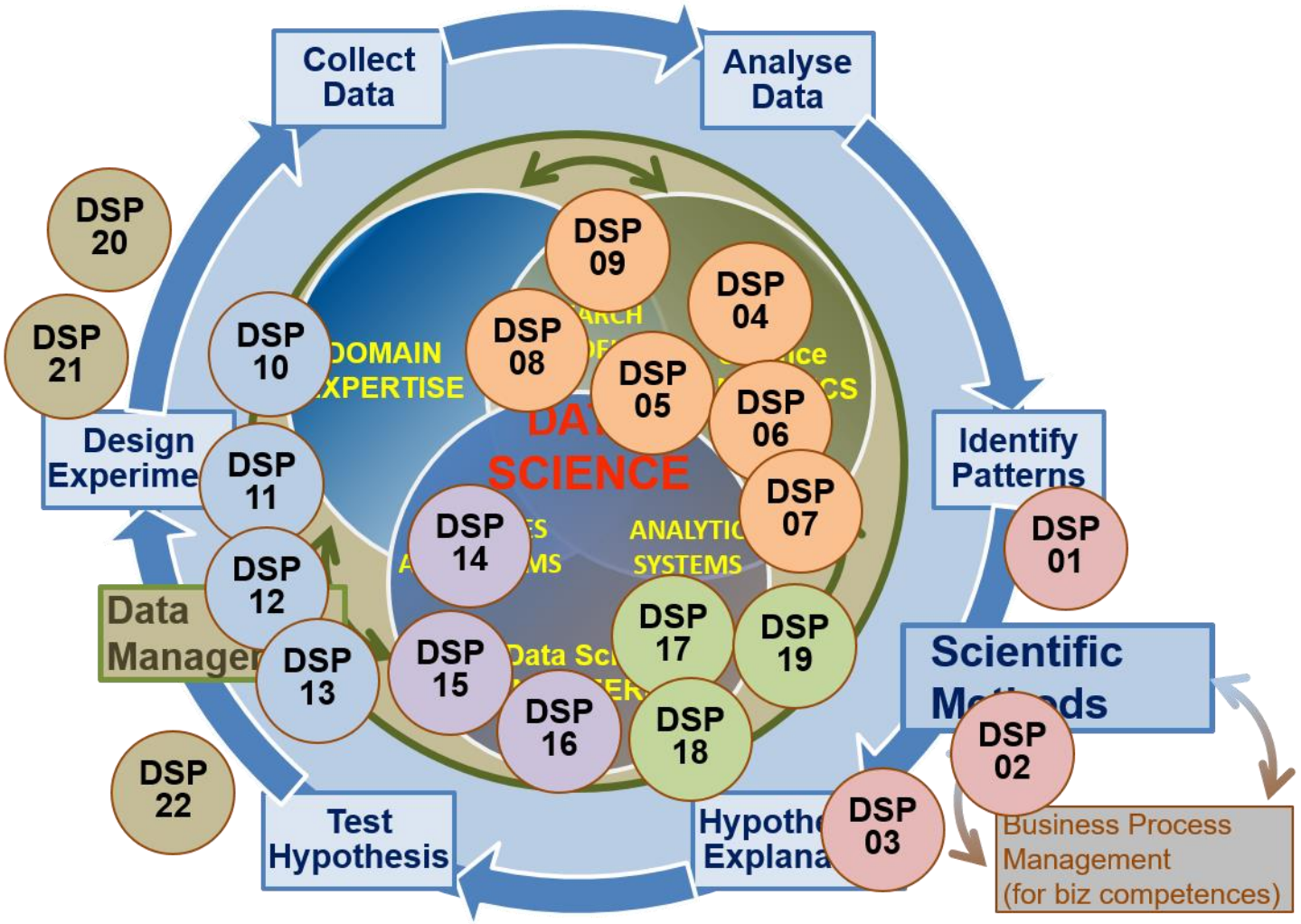
- Basis for the definition of the Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS)
  - CF-DS => Learning Outcomes (MC-DS) => Knowledge Areas (DS-BoK)
  - CF-DS => Data Science taxonomy of scientific subjects and vocabulary
- Data Science professional profiles definition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

Data Science Professions Family



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

CF-DS and Data Science Professional Profiles

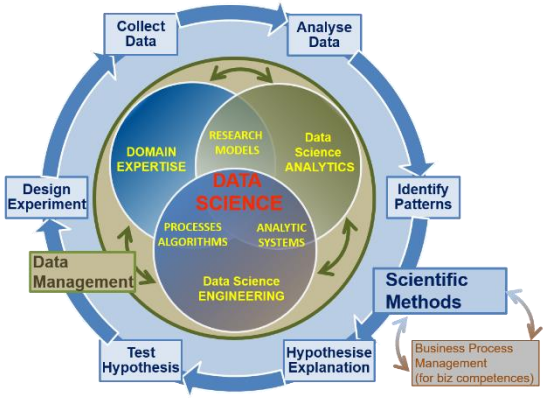


EDSF for Education and Training

- 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

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

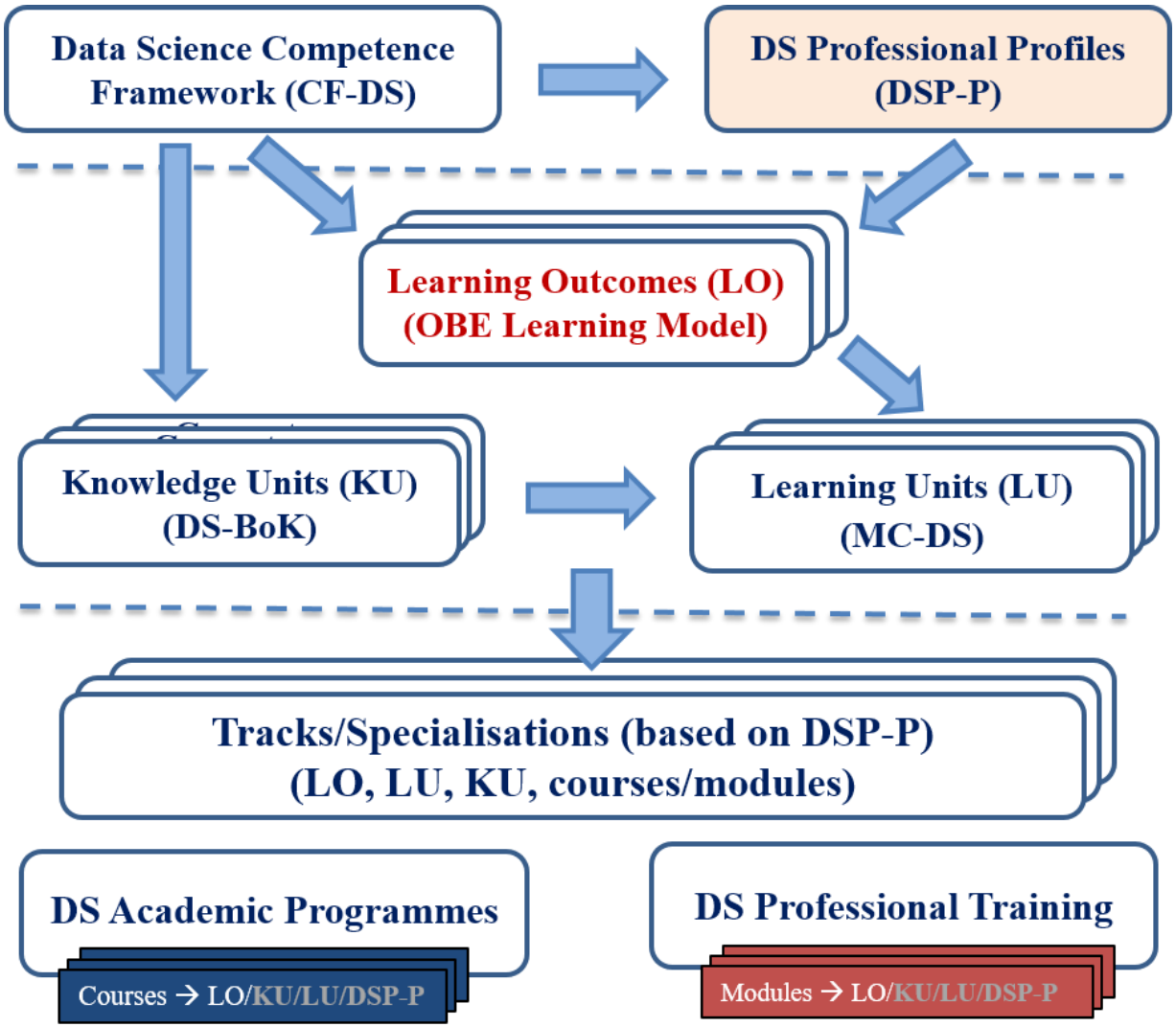
- Highlighted in red: Considered (Research) Data Management literacy (minimum required knowledge)

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)

Outcome Based Educations and Training Model



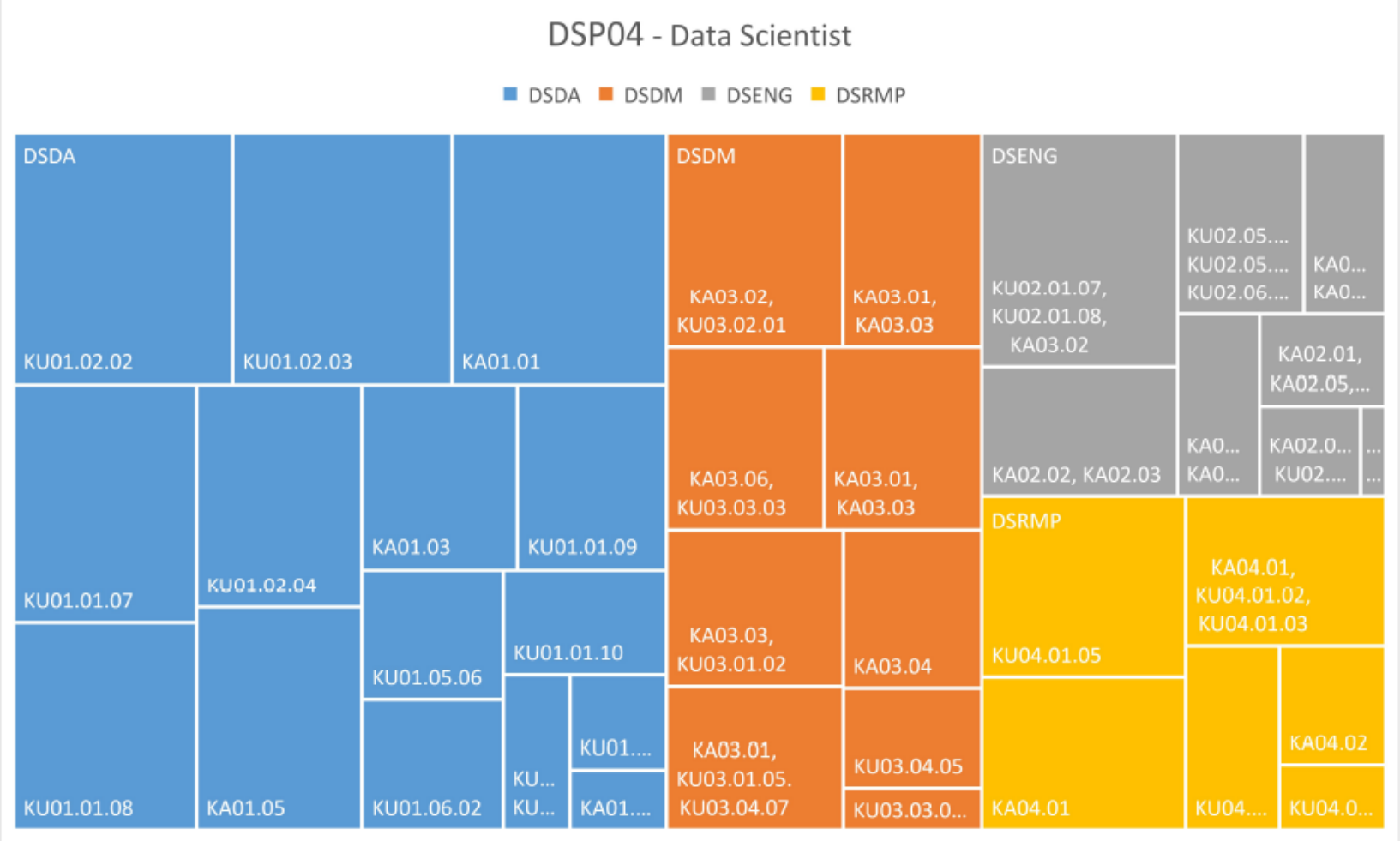
From Competences and DSP Profiles

to Learning Outcomes (LO) and

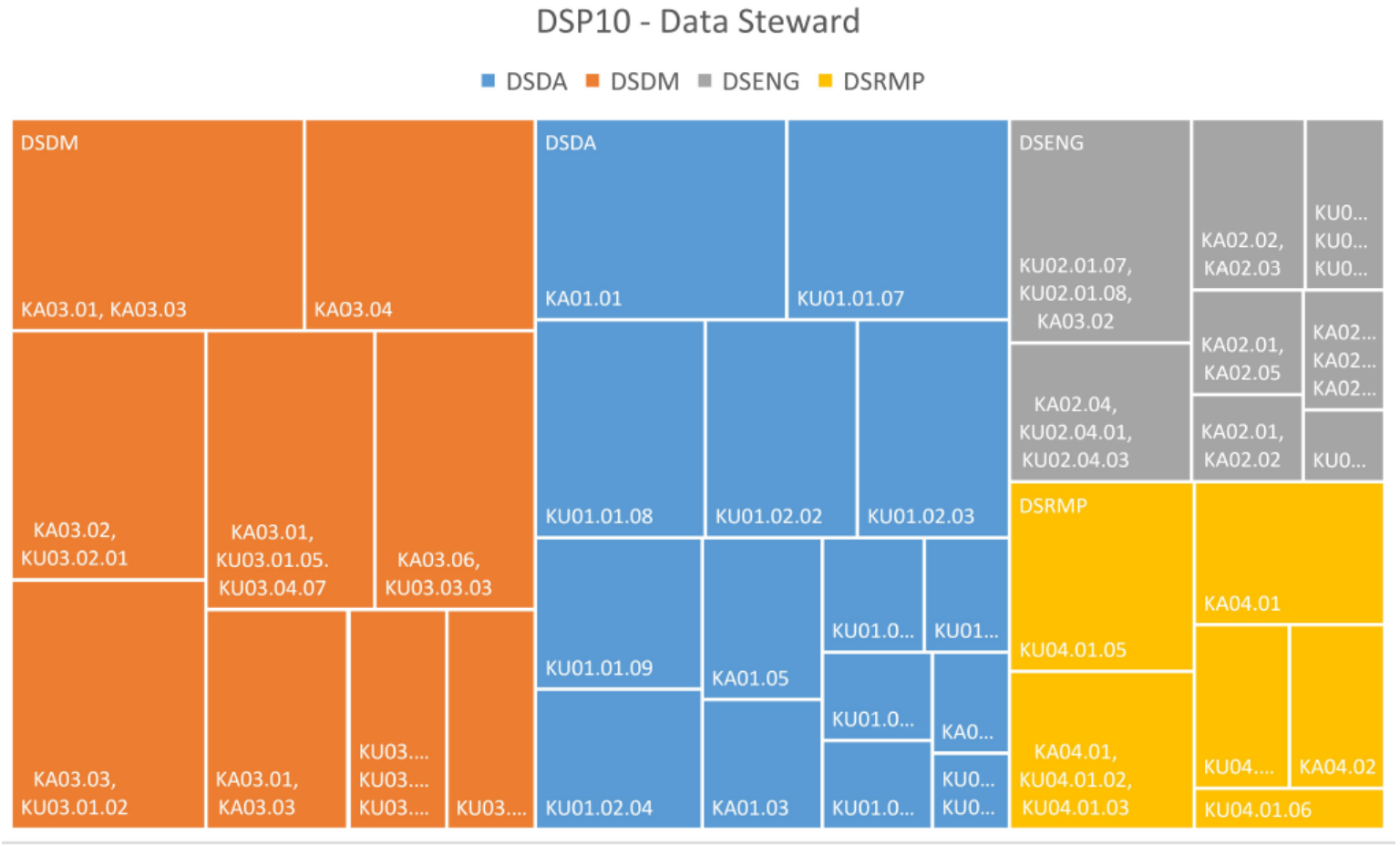
to Knowledge Unites (KU) and Learning Units (LU)

- EDSF allow for customized educational courses and training modules design

DSP04 – Data Scientist MC structure



DSP10 – Data Steward MC structure

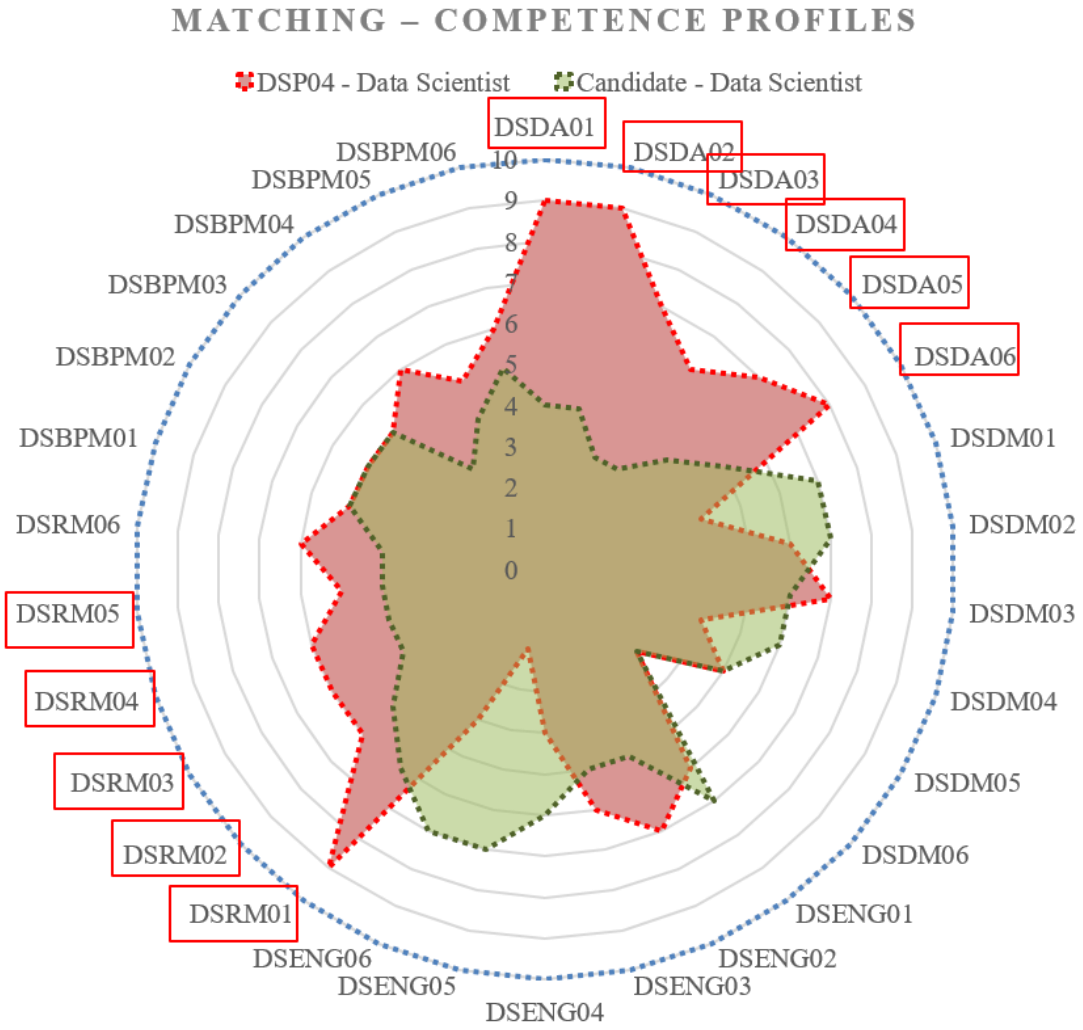


DSP04 Data Scientist – Required practical skills and Hands-on labs

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

Individual Competences Benchmarking

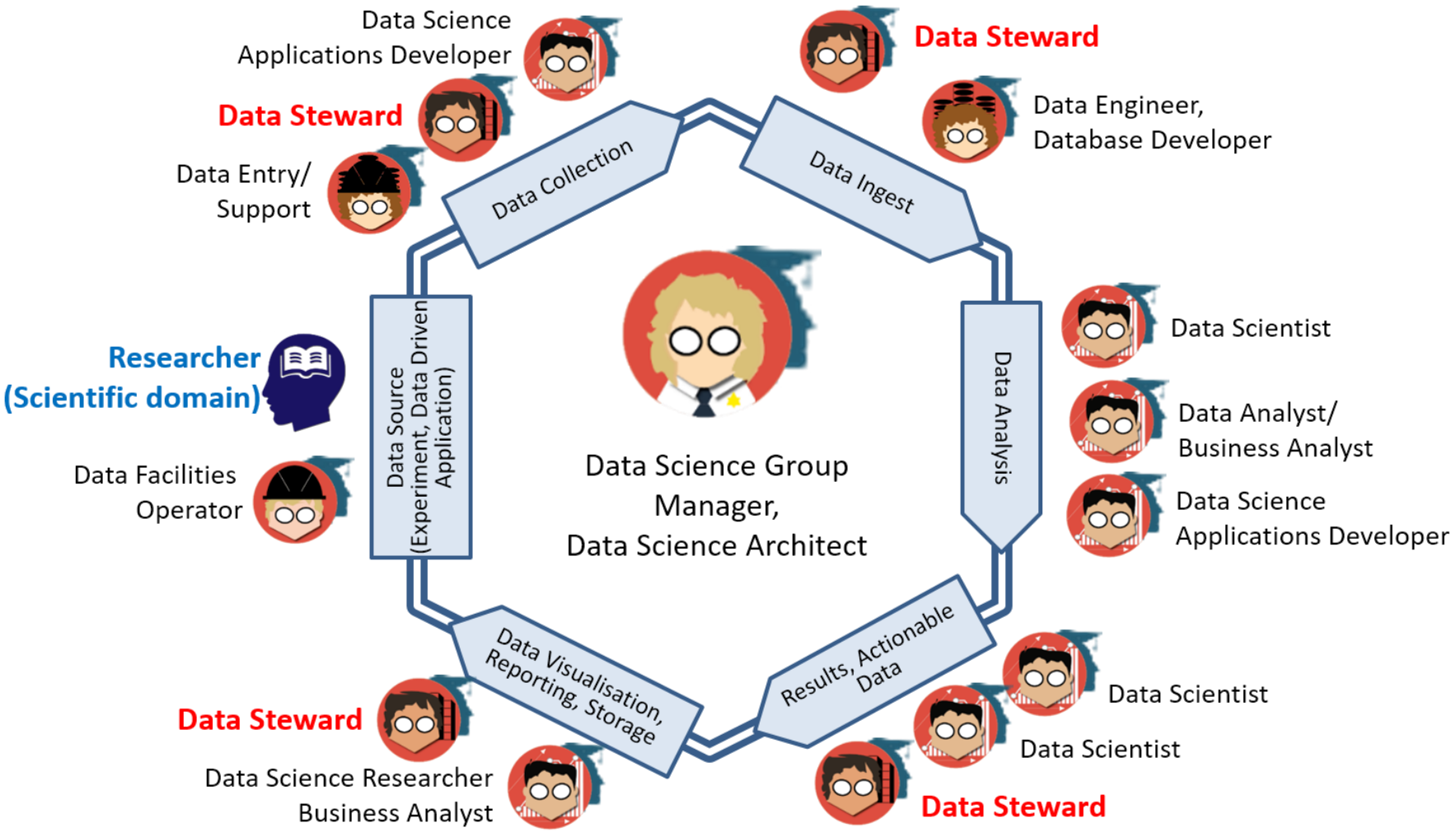


Individual Education/Training Path based on Competence benchmarking

- 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.

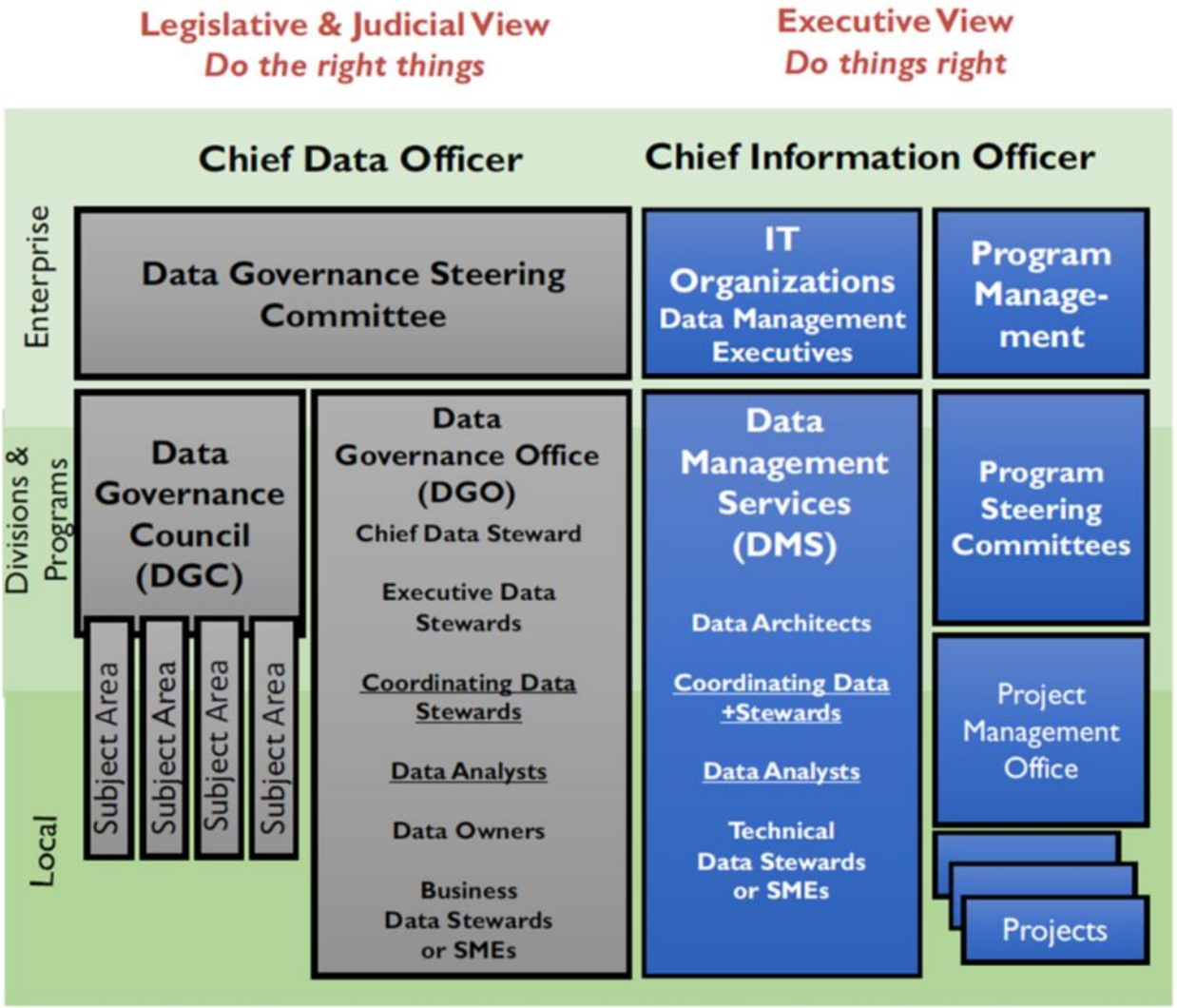
Growing role and demand for Data Stewards and data stewardship

Data Stewardship in Research and FAIR Principles

- FAIR Initiative by Dutch Techcentre for Life Science (DTLS) – Prof. Barend Mons
  - Supported by Germany, France, Spain, UK, USA
  - Part of Horizon 2020 Programme
- FAIR Principles for research data:
  Findable – Accessible – Interoperable - Reusable
- Data Stewards as a key bridging role between Data Scientists as (hard)core data experts and scientific domain researchers (HLEG EOSC report)
- Current definition of the Data Steward (part of Data Science Professional profiles)
  - Data Steward is a **data handling and management professional** whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation.
  - Data Steward creates data model for **domain specific data**, support and advice domain scientists/ researchers during the whole research cycle and data management lifecycle.

HLEG report on European Open Science Cloud (October 2016)

Data Governance Organisation Parts and Roles



[ref] DAMA-DMBOK Data Management Body of Knowledge, 2nd Edition, 2017

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

DAMA Data Management Body of Knowledge (DMBOK)

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.

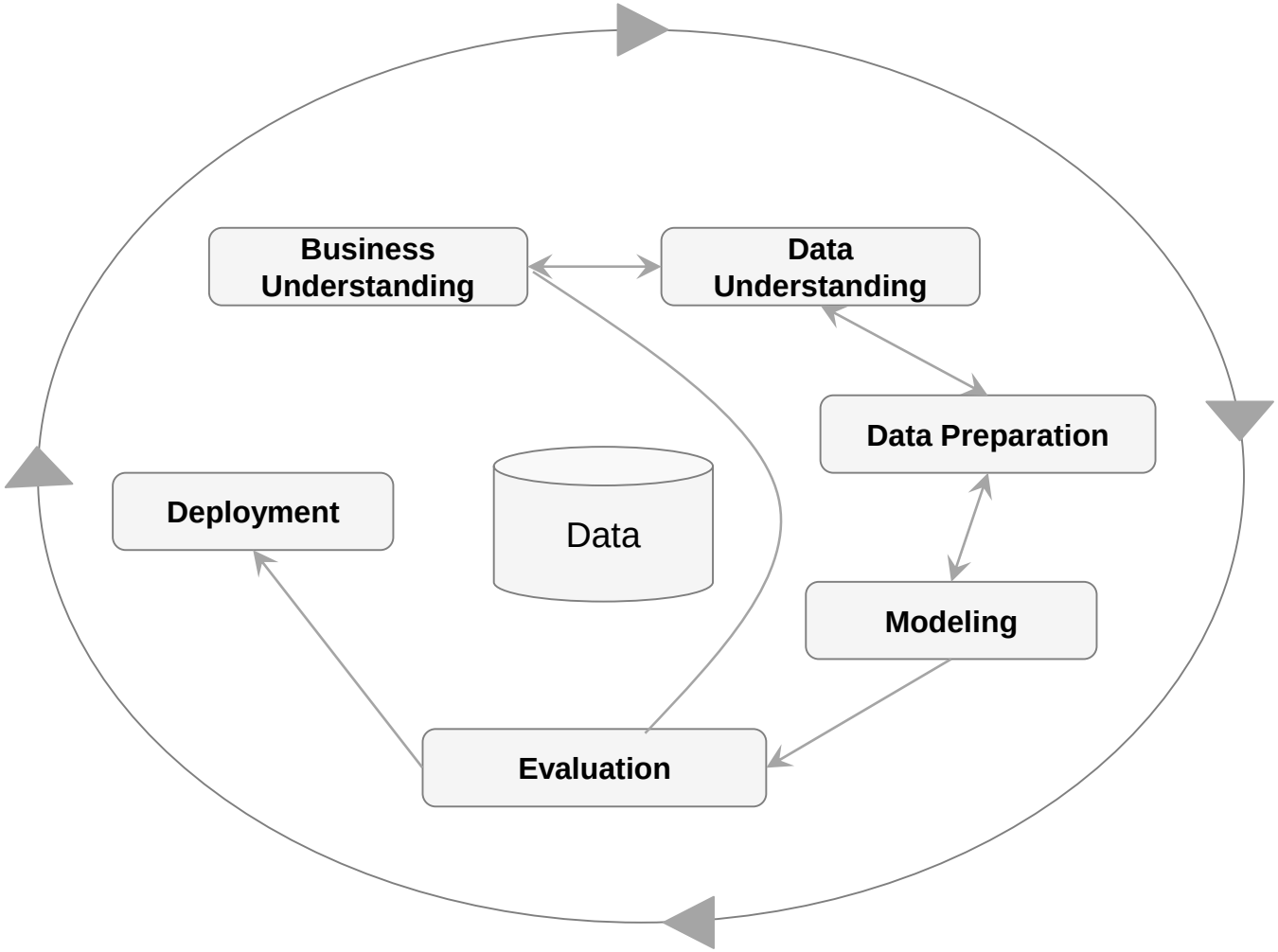
Discussion: How to become a Data Scientist

- 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

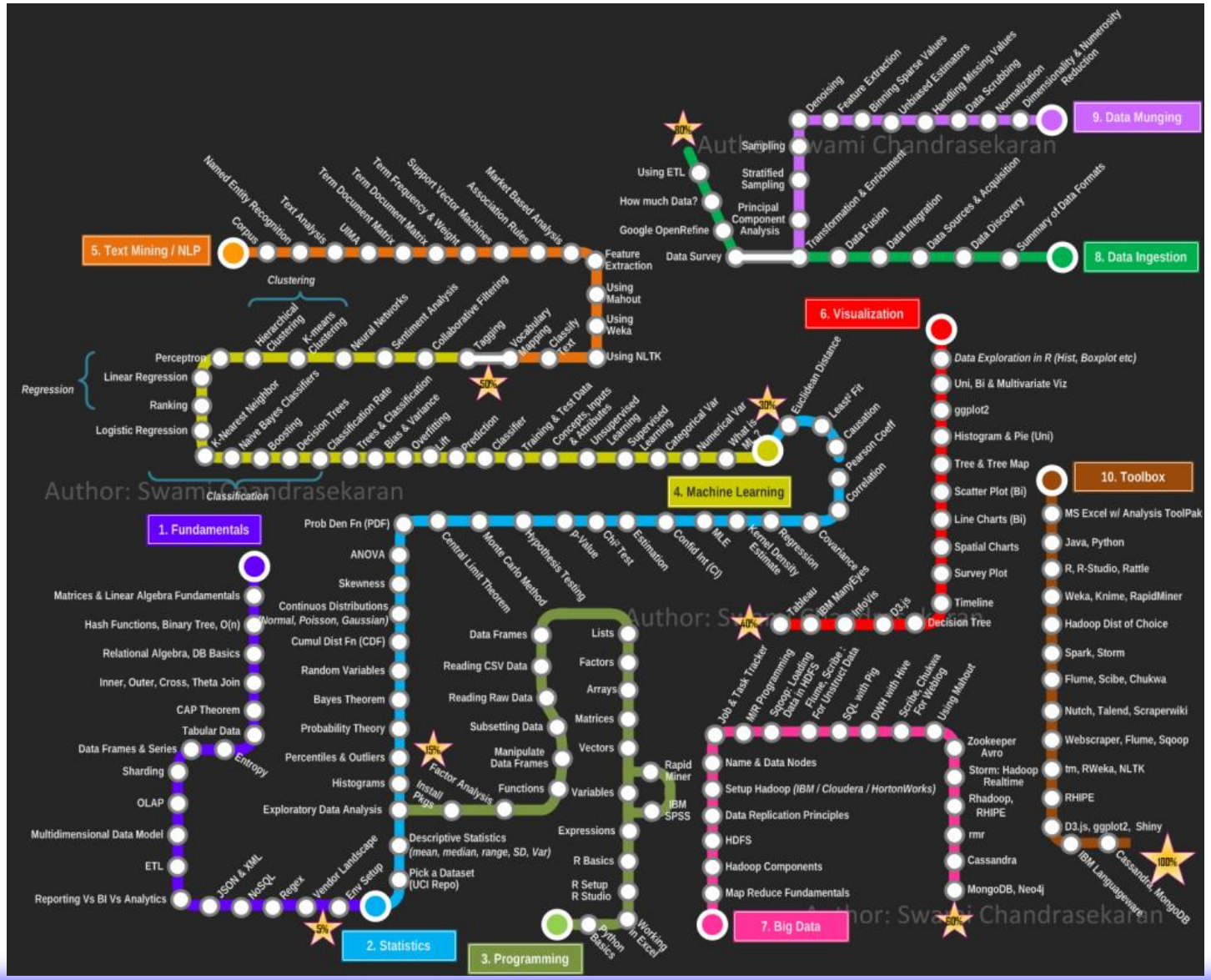
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis,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

CRISP DM process: Processes and Data Lifecycle



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

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 (MVA)
- (IBM – in transition)
- DataCamp
- Coursera, Udacity
- Certification and training PMI, DAMA, IIBA

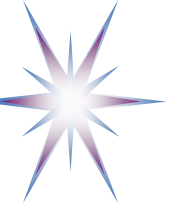
Open Data and Educational Datasets

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

Questions and discussion

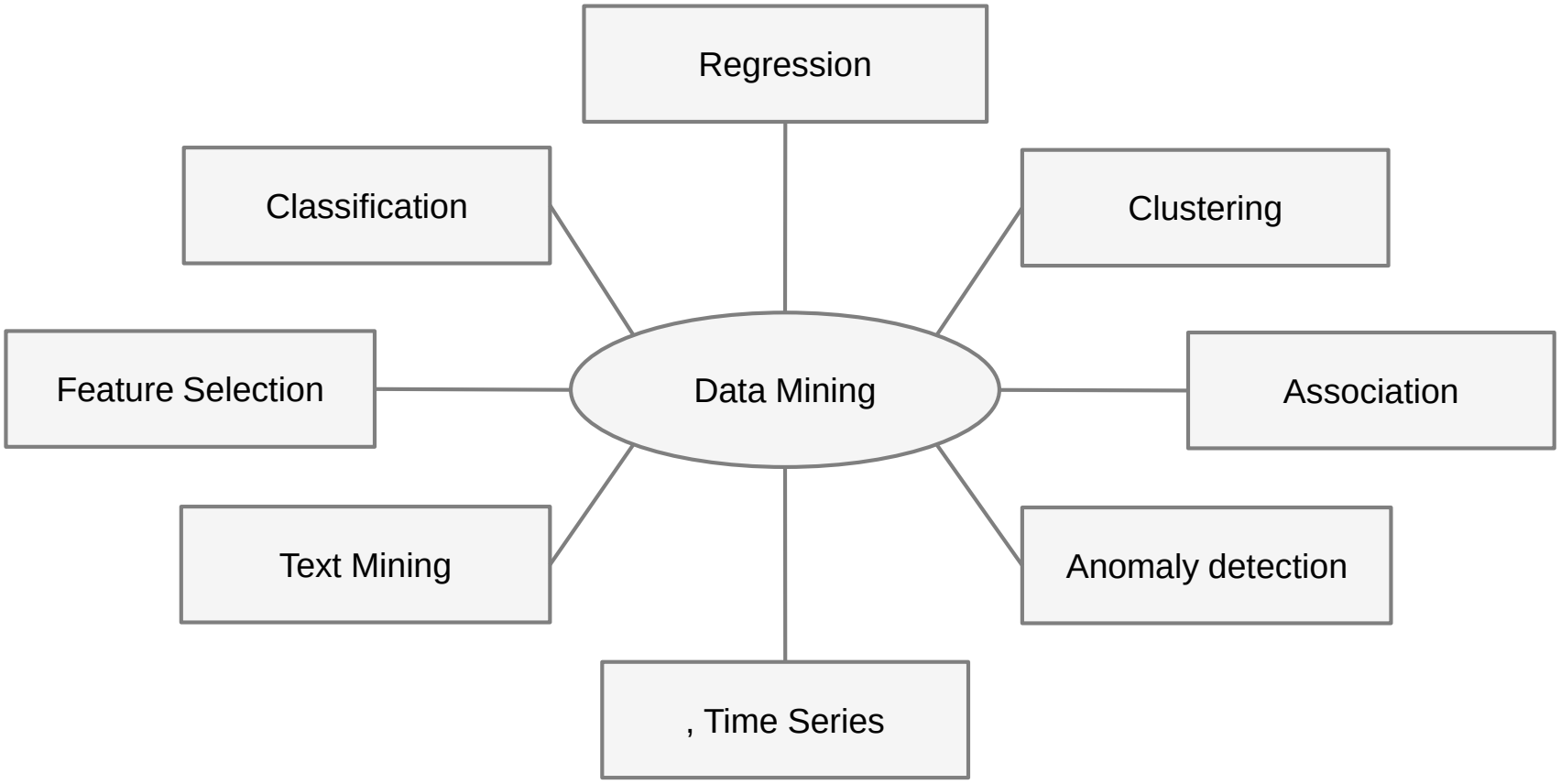
Other related links

- 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

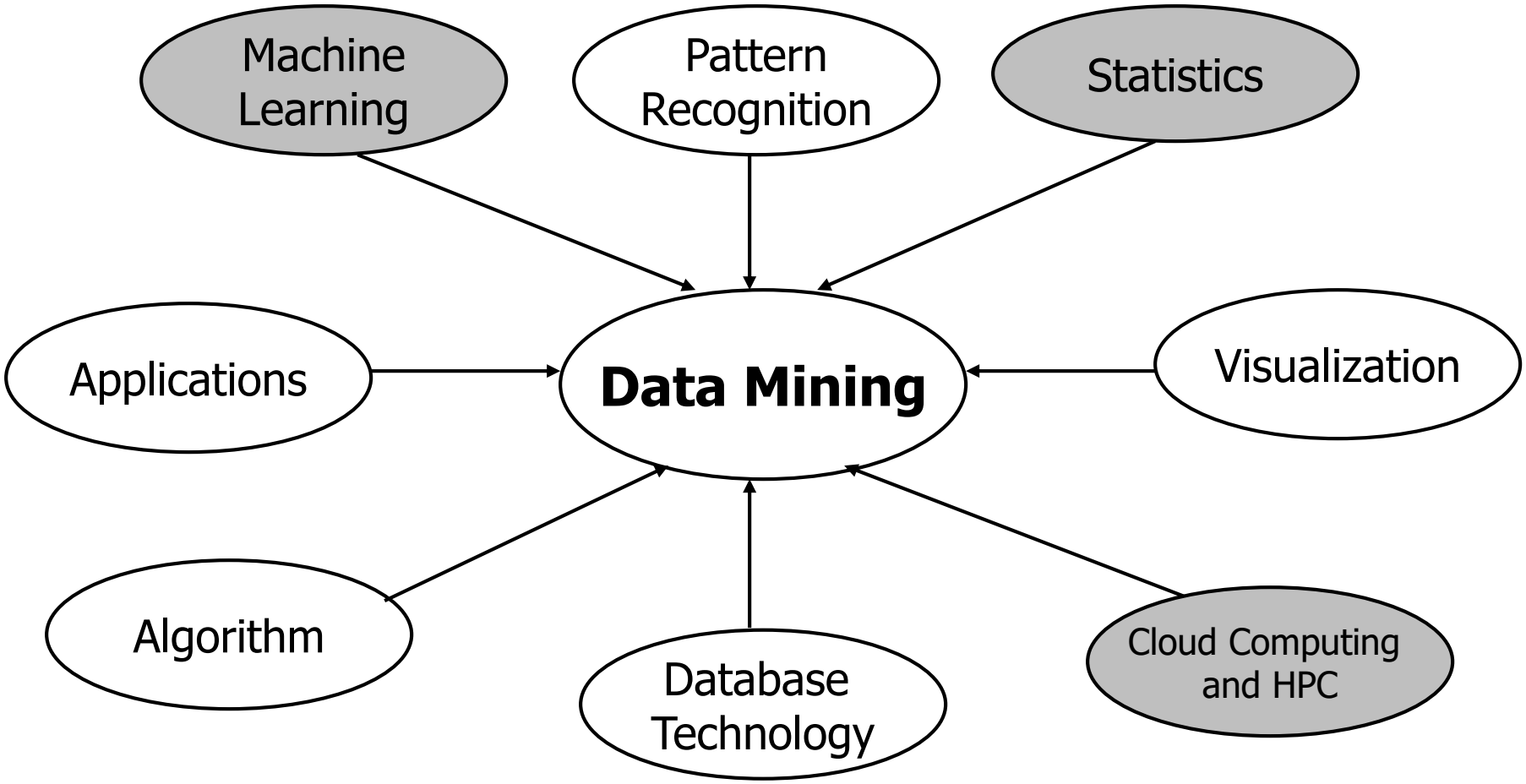


Additional materials

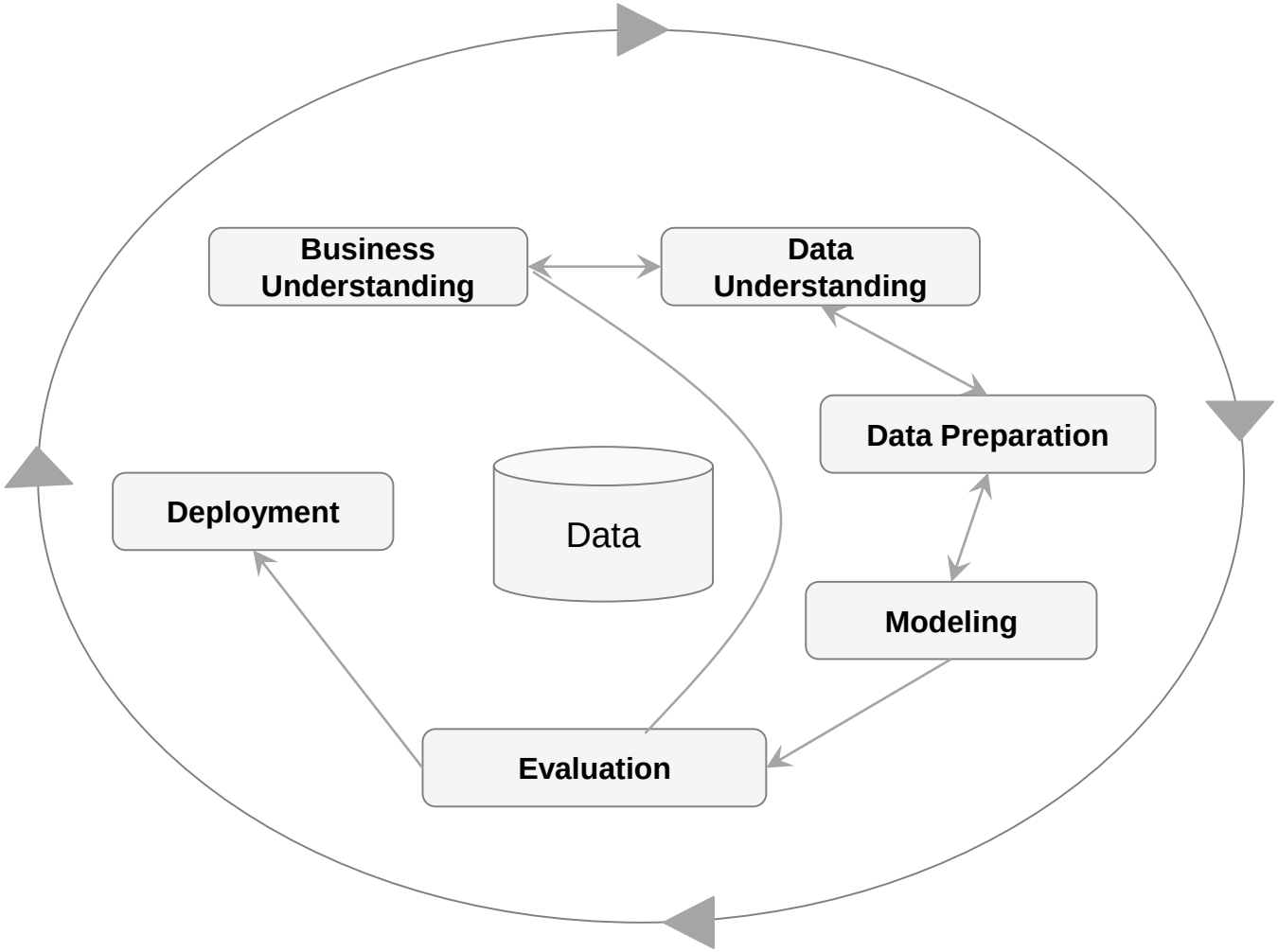
Types of Data Mining (branch of Data Analysis)



Data Mining: Confluence of Multiple Disciplines

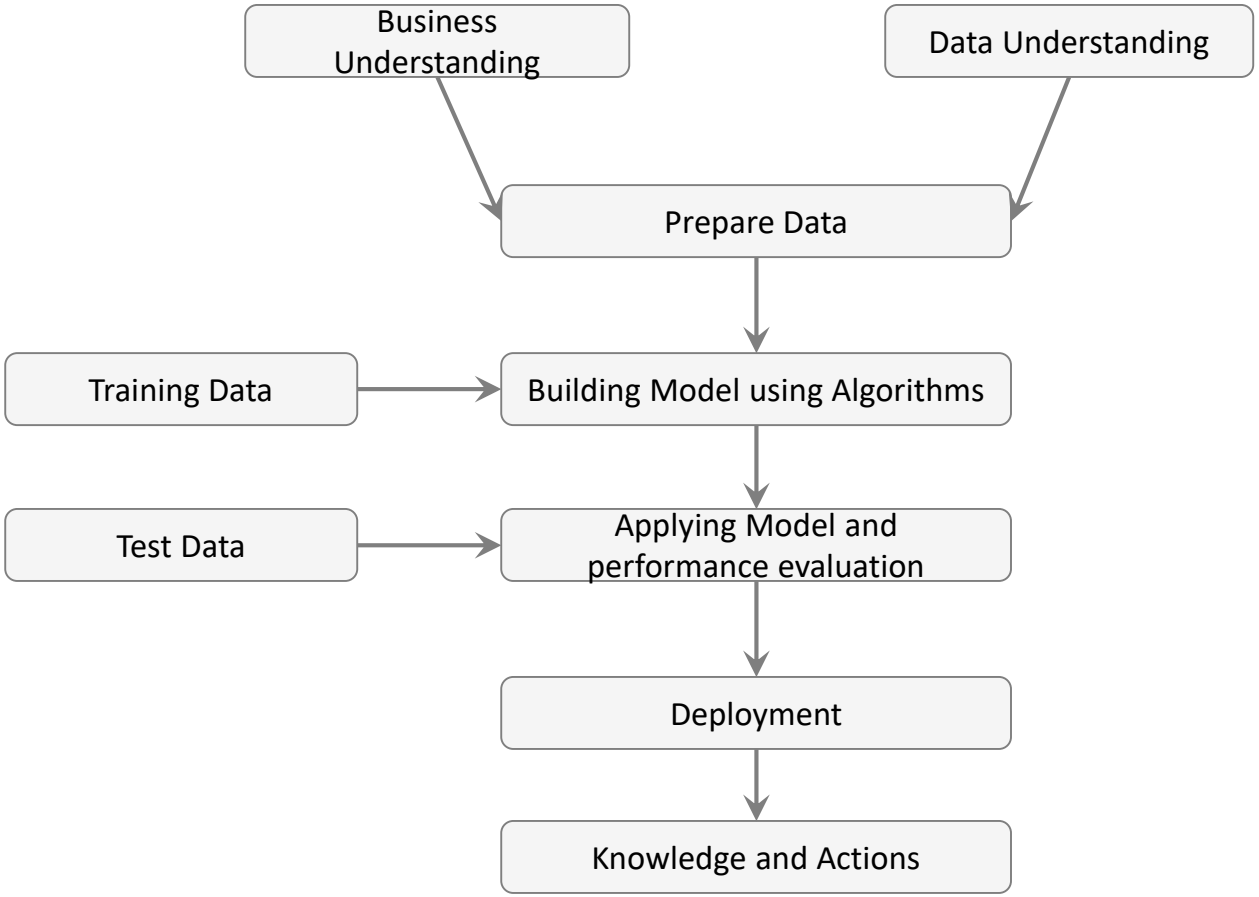


CRISP DM process: Processes and Data Lifecycle



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

Process of Data Analysis



1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge