The rise of Data Scientist

Dealing with Big Data in Agile Data Driven Organisations and New Organisational Roles

Yuri Demchenko, EDISON Project
University of Amsterdam

Big Data Innovation Conference BDIC2016

16-17 November 2016, Frankfurt, Germany

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

Grant 675419 (INFRASUPP-4-2015: CSA)

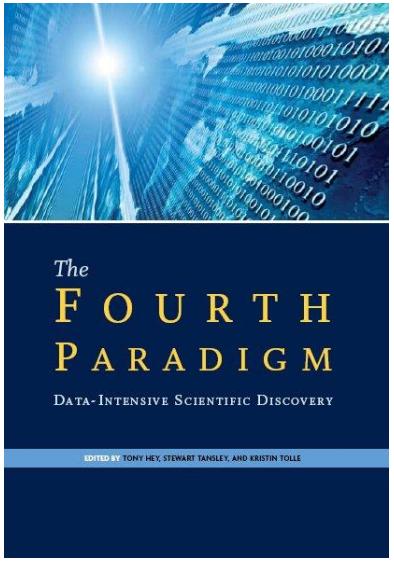
Outline

- Background and motivation
  - Technologies timeline and data driven technologies
  - New type of Agile Data Driven Organisations (ADDO)
  - Demand for Data Science and data related professions
  - European initiatives related to Digital Single Market and Digital Skills Agenda
- EDISON Data Science Framework (EDSF)
  - From Data Science Competences to Body of Knowledge and Model Curriculum
- Data Science Competence Framework: Essential competences and skills
- New organisational role and Data Science Professional profiles
- Educational and training aspects in Data Science and organisational roles
  - Example Research Data Management Literacy curriculum
  - Professional Certification and Self-made Data Scientists
- Wide spectrum of activities and initiatives worldwide to establish Data (Science) professions family

Technologies Timeline

- **Service Oriented Architecture (SOA)**: First proposed in 1996 and revived with the Web Services advent in 2001-2002
  - Currently standard for industry, and widely used
- **Computer Grids**: Initially proposed in 1998 and finally shaped in 2003 with the Open Grid Services Architecture (OGSA) by Open Grid Forum (OGF)
  - Migrated to cloud platform, serves as collaborative environment for research
- **Cloud Computing**: Initially defined in 2008, matured 2011 *– Now in productive phase*
  - Defined ***new features, capabilities, operational/usage models*** and actually provided a guidance for the new technology development
- **Big Data and Data Intensive Science**: Shaped in 2014 *- Growing adoption and maturity*
  - Defined as *Ecosystem* where data are the main driving component
  - Critical need for *new digital skills and professions*
  - Recognised need for the Data Science definition as new scientific discipline and and Data Scientist as a new profession
- **Cognitive Technologies and AI**: Emerged 2016 *- Gaining recognition and finding use*
  - Leverages and consolidation of recent foundational technologies and raised role of large scale computing

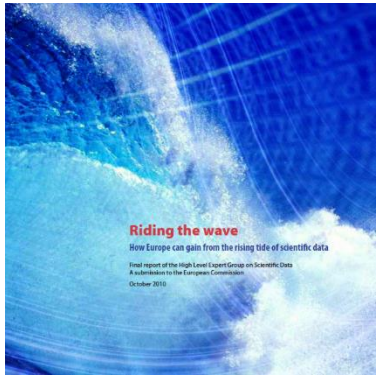
Visionaries and Drivers: Seminal works, High level reports, Activities



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/

NIST Big Data Working Group (NBD-WG)
http://bigdatawg.nist.gov/

ISO/IEC JTC1 Big Data Study Group (SGBD)
http://jtc1bigdatasg.nist.gov/home.php

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

The Fourth Paradigm of Scientific Research

1. Theory and logical reasoning
2. Observation or Experiment
   - E.g. Newton observed apples falling to design his theory of mechanics
   - But 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 beats hypnotized theory

Agile Data Driven Organisation/Enterprise (ADDO/ADDE)

Four pillars of the Agile Data Driven Organisation (ADDO)

- Enterprise technological agility/mobility
- Event driven analytics, cognitive technologies
- Using new source of data to correlate with the company's business (location, cultural, social media, sentiments, profiling and targeting)
- Cloud as a general computing and data management infrastructure

Steps/conditions to become ADDO

- Curious thinking to explore new data use, including use of new data sources
- Management must be receptive to the role of Big Data analytics and possible recommendations as outcome of new discovered factors and relations
- Key/driving role of the ***Data Scientist***
  - New emerging profession capable of using scientific research methods
- Employees must understand that Data Driven approach will help them doing their work
- Availability of data analytics tools to support and facilitate data self-service by employees
  - Tools and applications must provide overview of the process and properties of the data collected
- Comprehensive and customizable data visualisation
- Incorporating mobile platform for data collection and data delivery

What's new in "data-driven" ADDE

- Data Driven methods and model have been used long time ago – What is new now?
- IBM WoW 2016 Conference (23-24 Oct 2016 Las Vegas) discussions
  - Rise of cognitive technologies and new understanding of "data driven"
- All analytics reports at/by IBM are required to be correlated with "modern sources of data"
  - Social Network/Media (human&incentives): Twitter, facebook, Linkedin, Pinterest, Slack, YouTube, Google Plus+, Tumblr, Instagram, Flikr, etc
  - Additionally (event&trends): news, weather, biodiversity, fashion trends, etc.
- Cloud based cognitive platforms to provide global services for data sources alignment

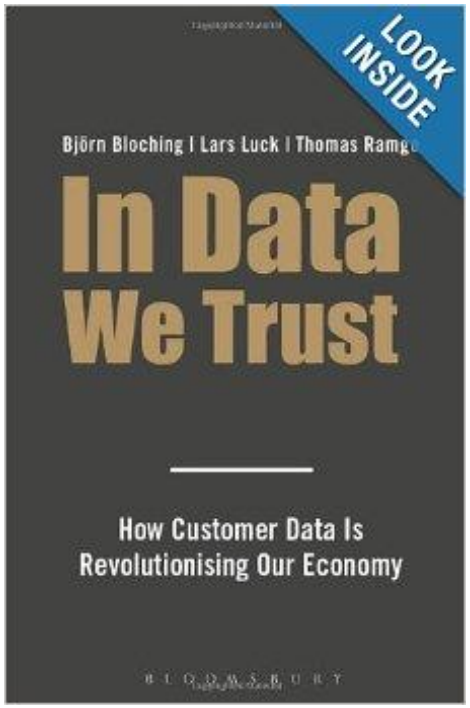
Data Driven Victories and Failures - Politics

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 Big Data technologies
- US Election 2016
  - Clinton's campaign – "Data driven" but using only upper layer of 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."

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)

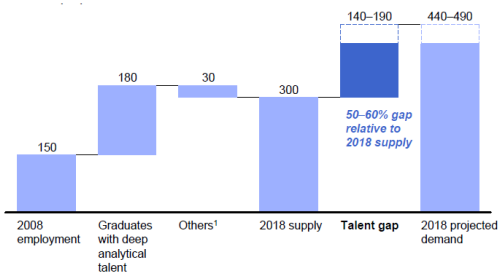
- A strategy for tomorrow's data world

Data-Driven Brexit: A Wakeup Call for Analysts
By Barry Devlin, June 28, 2016

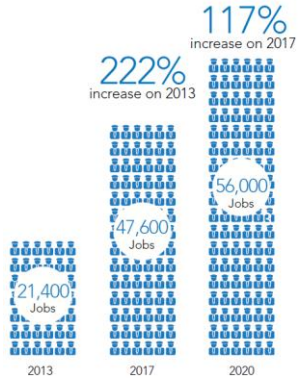
- Article “In Data we trust” by T.Edsall in The New York Times
- Multimillion-dollar contract for data management and collection services awarded May 1, 2013 to Liberty Work to build advanced list of voters

- 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.

Demand for Data Science and data related professions



- McKinsey Global Institute on Big Data Jobs (2011) http://www.mckinsey.com/mgi/publications/big_data/index.asp
  - Estimated gap of 140,000 - 190,000 data analytics skills by 2018
- UK Big Data skills report 2014
  - 6400 UK organisations with 100+ staff will have implemented Big Data Analytics by 2020
  - Increase of Big Data jobs from 21,400 (2013) to 56,000 (2017)



- IDC Report on European Data Market (2015)
  - Number of data workers 6.1 mln (2014) - increase 5.7% from 2013
  - Average number of data workers per company 9.5 - increase 4.4%
  - Gap between demand and supply 509,000 (2014) or 7.5%
- HLEG report on European Open Science Cloud (October 2016) identified need for data experts and data stewards
  - Recommendation: Allocate 5% grant funding for Data management and preservation
  - Estimation: More than 80,000 data stewards (1 per every 20 scientists)
  - Core data experts need to be trained and their career perspective improved

Recent European Commission Initiatives 2016

Digitising European Industry: Reaping the full benefits of a **Digital Single Market**. COM(2016) 180 final, Brussels, 19.4.2016

- The need for new multidisciplinary and digital skills in particular Data Scientist
  - Expected rapidly growing demand will lead to more than 800 000 unfilled vacancies by 2020

European Cloud Initiative - Building a competitive data and knowledge economy in Europe, COM(2016) 178 final, Brussels, 19.4.2016

- **European Open Science Cloud (EOSC)** and European digital research and data infrastructure
  - To offer 1.7 million European researchers and 70 million professionals in science and technology open and seamless services for ***storage, management, analysis and re-use*** of research data
- Address growing demand and shortage of data-related skills

A New Skills Agenda for Europe, COM(2016) 381 final Brussels, 10.6.2016

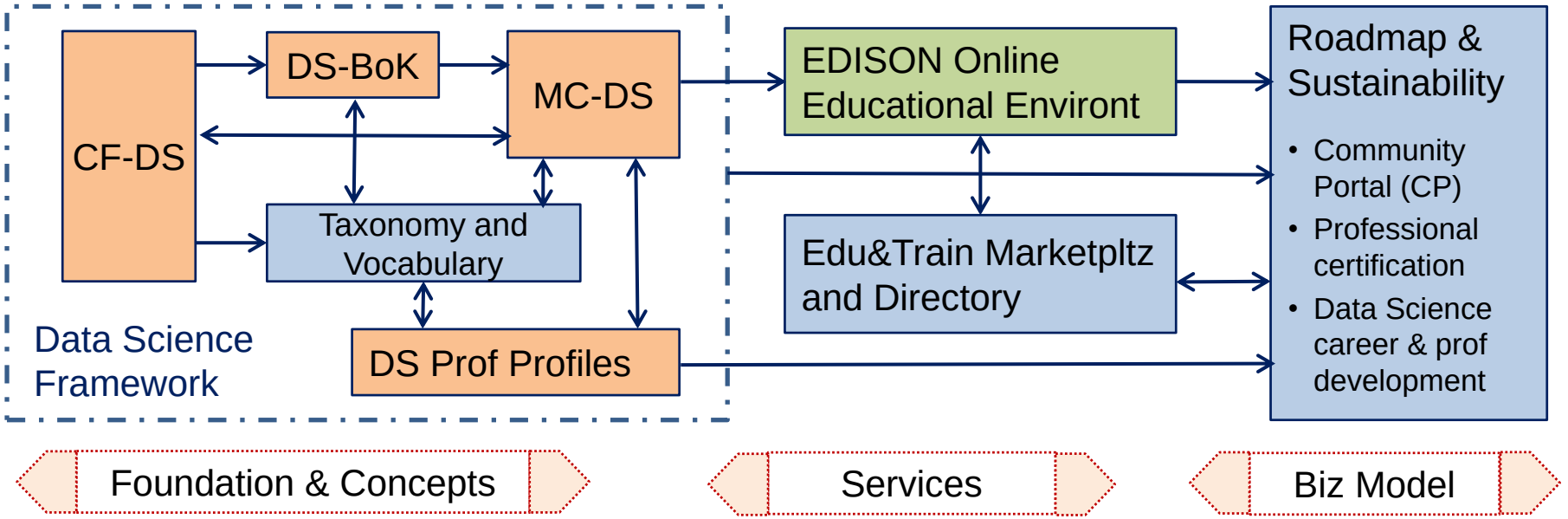
- Addresses the need for digital and complementary skills, ensure young talents flow into data driven research and industry
- Launch **Digital Skills and Jobs Coalition** (1st December 2016, Brussels) to develop comprehensive national digital skills strategies by mid-2017

OECD on Digital Economy

- 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

EDISON Data Science Framework (EDSF) Release 1 (October 2016)



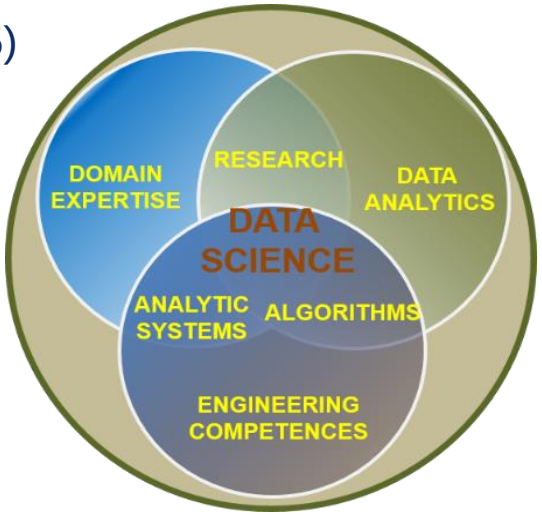
- EDISON Framework components
  - CF-DS – Data Science Competence Framework
  - DS-BoK – Data Science Body of Knowledge
  - MC-DS – Data Science Model Curriculum
  - DSP – Data Science Professional profiles
  - Data Science Taxonomies and Scientific Disciplines Classification
  - EOEE - EDISON Online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

Data Scientist definition

Based on the definitions by NIST Big Data WG (NIST SP1500 - 2015)

- *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 expected scientific and business value to science or industry*

- *Other definitions to admit such features as*
  - Ability to solve variety of business problems
  - Optimize performance and suggest new services for the organisation
  - Develop a special mindset and be statistically minded, ***understand raw data** and **"appreciate data as a first class product"***



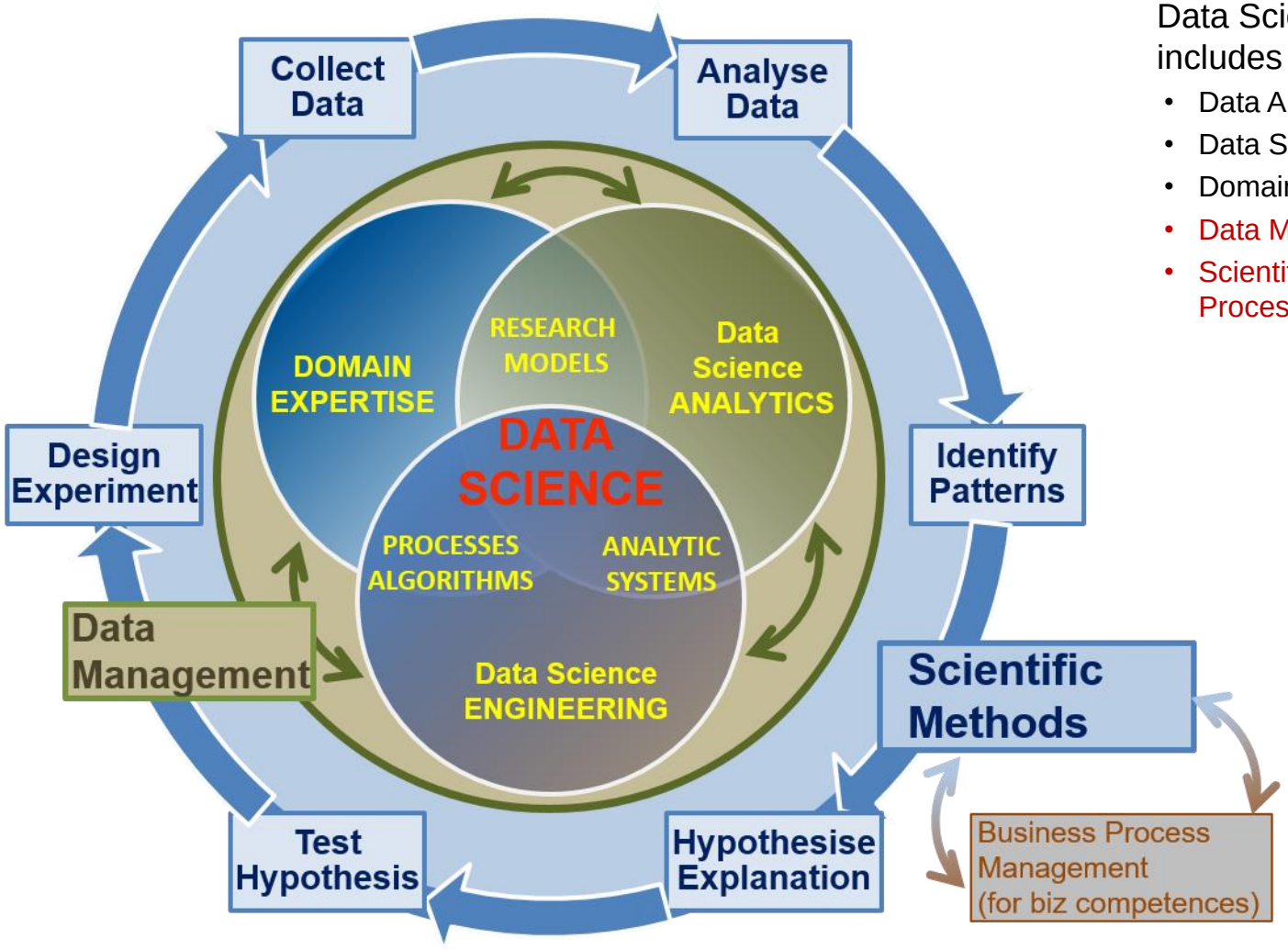
[ref] Legacy: NIST BDWG definition of Data Science

- ***Data science** is the empirical synthesis of actionable knowledge and technologies required to handle data from raw data through the complete data lifecycle process.*
- ***Big Data** is the technology to build system and infrastructures to process large volume of structurally complex data in a time effective way*

Identified Data Science Competence Groups

- Commonly accepted Data Science competences/skills groups include
  - Data Analytics or Business Analytics or Machine Learning
  - Engineering or Programming
  - Subject/Scientific Domain Knowledge
- EDISON identified 2 additional competence groups demanded by organisations
  - **Data Management, Curation, Preservation**
  - **Scientific or Research Methods and/vs Business Processes/Operations**
- Other skills commonly recognized aka “soft skills” or “social or professional intelligence”
  - Inter-personal skills and team work, cooperativeness
- All groups need to be represented in Data Science curriculum and training programmes
  - Challenging task for Data Science education and training
- Another aspect of integrating Data Scientist into organisation structure
  - General Data Science (and Data) **literacy** for all involved roles and management
  - Common agreed and understandable way of communication and **information/data presentation**
  - ***Role of Data Scientist: Provide a kind of literacy advice and guidance to organisation***

Data Science Competence Groups - Research



Data Science Competence includes 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- Data Management
- Scientific Methods (or Business Process Management)

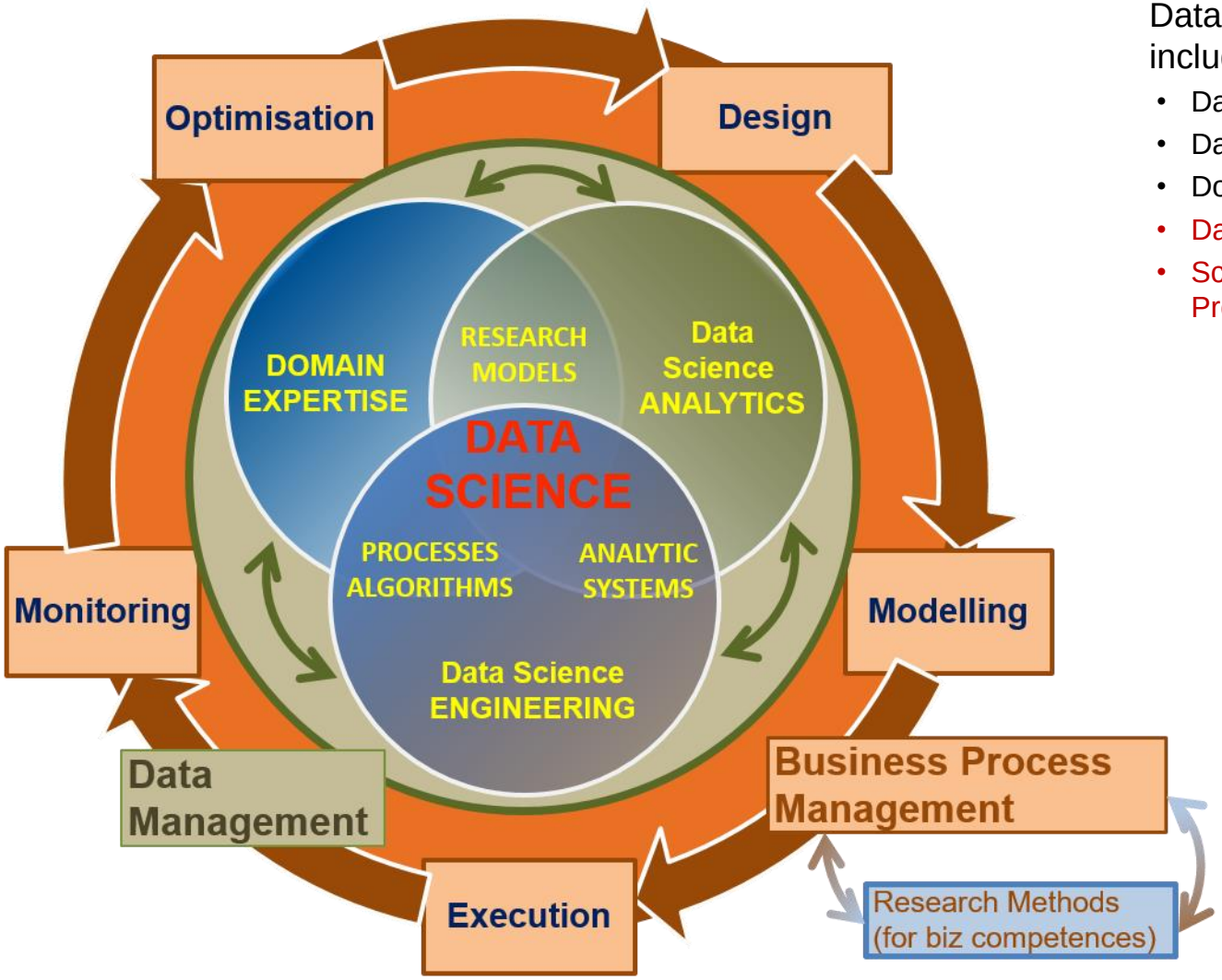
Scientific Methods

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

Business Operations

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

Data Science Competences Groups – Business



Data Science Competence includes 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- Data Management
- Scientific Methods (or Business Process Management)

Scientific Methods

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

Business Process Operations/Stages

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

Identified Data Science Competence Groups

	Data Science Analytics (DSDA)	Data Management (DSDM)	Data Science Engineering (DSENG)	Research/Scientific Methods (DSRM)	Data Science Domain Knowledge, e.g. Business Processes (DSDK/DSBPM)
0	Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management	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	Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
1	**DSDA01** Use predictive analytics to analyse big data and discover new relations	**DSDM01** Develop and implement data strategy, in particular, Data Management Plan (DMP)	**DSENG01** Use engineering principles to design, prototype data analytics applications, or develop instruments, systems	**DSRM01** Create new understandings and capabilities by using scientific/ research methods or similar domain related development methods	**DSBPM01** Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
2	**DSDA02** Use statistical techniq to deliver insights	**DSDM02** Develop data models including metadata	**DSENG02** Develop and apply computational solutions	**DSRM02** Direct systematic study toward a fuller knowledge or understanding of the observable facts	**DSBPM02** Participate strategically and tactically in financial decisions
3	**DSDA03** Develop specialized …	**DSDM03** Collect integrate data	**DSENG03** Develops specialized tools	**DSRM03** Undertakes creative work	**DSBPM03** Provides support services to other
4	**DSDA04** Analyze complex data	**DSDM04** Maintain repository	**DSENG04** Design, build, operate	**DSRM04** Translate strategies into actions	**DSBPM04** Analyse data for marketing
5	**DSDA05** Use different analytics	**DSDM05** Visualise cmplx data	**DSENG05** Secure and reliable data	**DSRM05** Contribute to organizational goals	**DSBPM05** Analyse optimise customer relatio

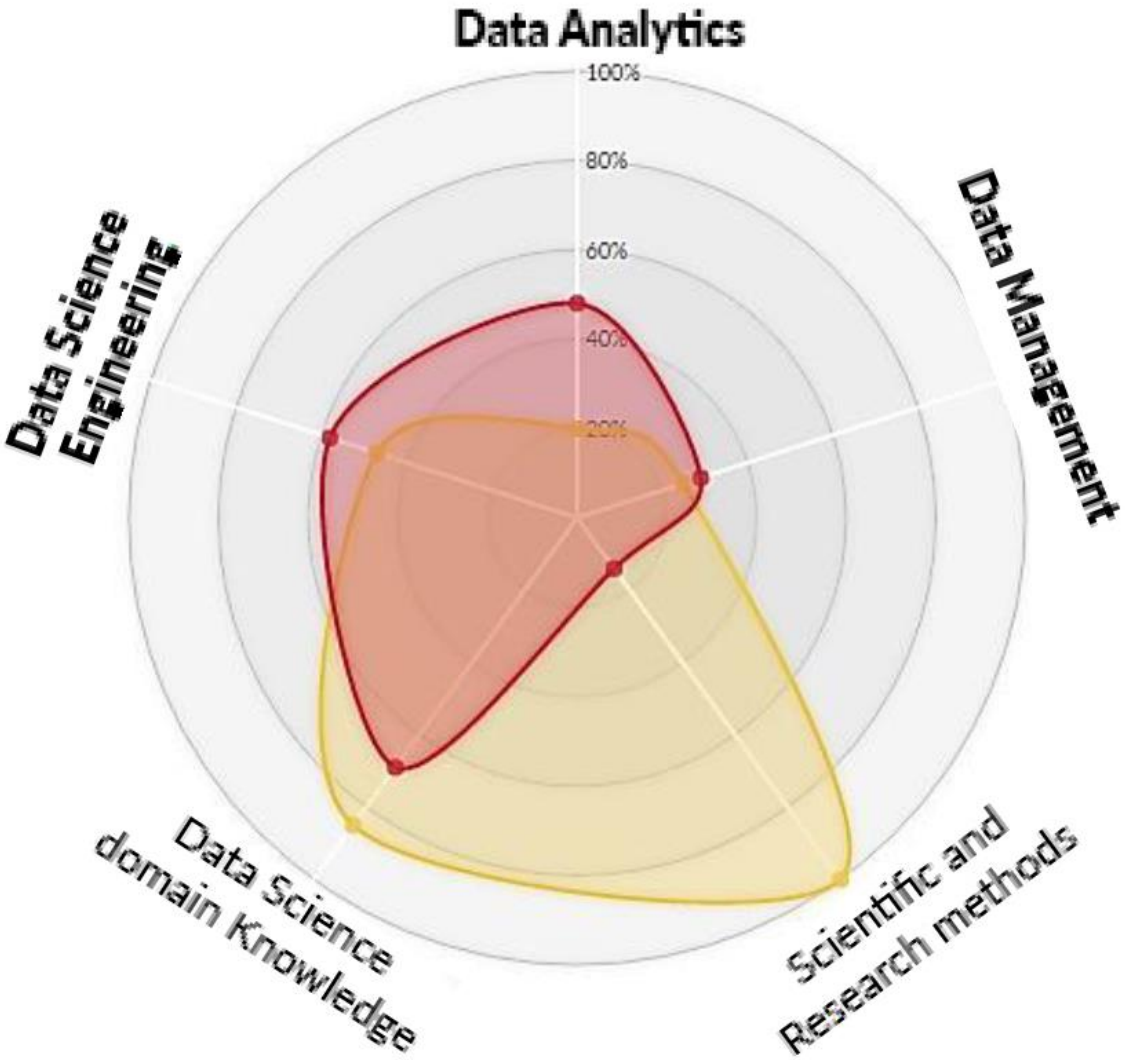
Identified Data Science *Skills/Experience* Groups

- **Group 1: Skills/experience related to competences**
  - Data Analytics and Machine Learning
  - Data Management/Curation (including both general data management and scientific data management)
  - Data Science Engineering (hardware and software) skills
  - Scientific/Research Methods or Business Process Management
  - Application/subject domain related (research or business)
  - Mathematics and Statistics
- **Group 2: Big Data (Data Science) tools and platforms**
  - Big Data Analytics platforms
  - Mathematics & Statistics applications & tools
  - Databases (SQL and NoSQL)
  - Data Management and Curation platform
  - Data and applications visualisation
  - *Cloud based platforms and tools*
- **Group 3: Programming and programming languages and IDE**
  - General and specialized development platforms for data analysis and statistics
- **Group 4: Soft skills or Social Intelligence**
  - Personal, inter-personal communication, team work, professional network

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

Example competence benchmark result



- Red polygon indicates the chosen profession
- Yellow polygon indicates the practitioner
- **Insufficient competences are shown in a bold typeface.**

[ref] Kim Hee et al (Frankfurt Univ) "Tailored Data Science Education using Gamification" - DTW2016 Workshop research paper

Data Science Professions Family

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, Digital 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

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

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

- Group Manager
- Data Science Architect
- Data Analyst
- Data Science Application programmer
- Data Infrastructure/facilities administrator/operator: storage, cloud, computation
- Data stewards

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

Data Science or Data Management Group/Department

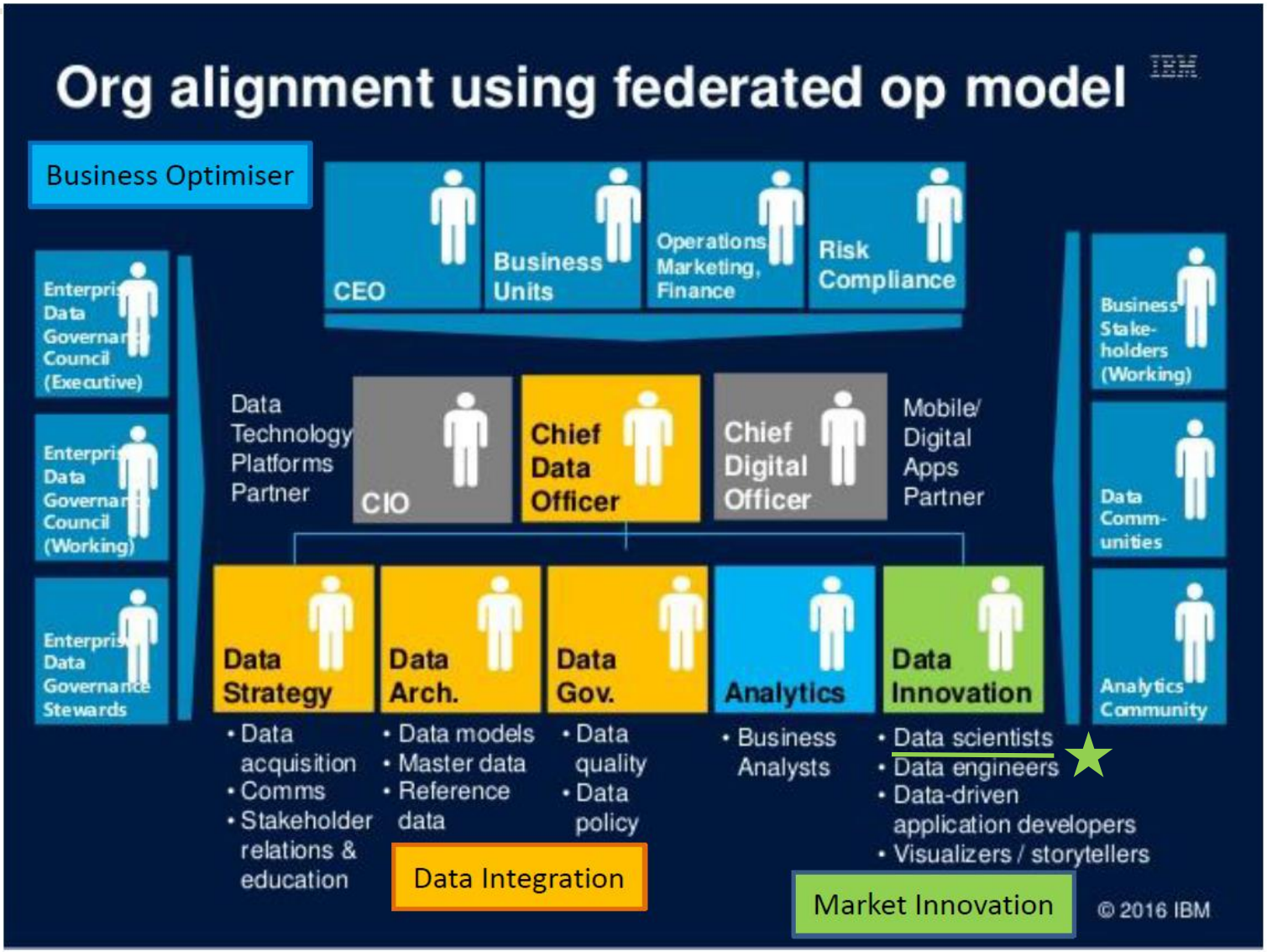
>> Reporting to CDO/CTO/CEO

- Providing cross-organizational services

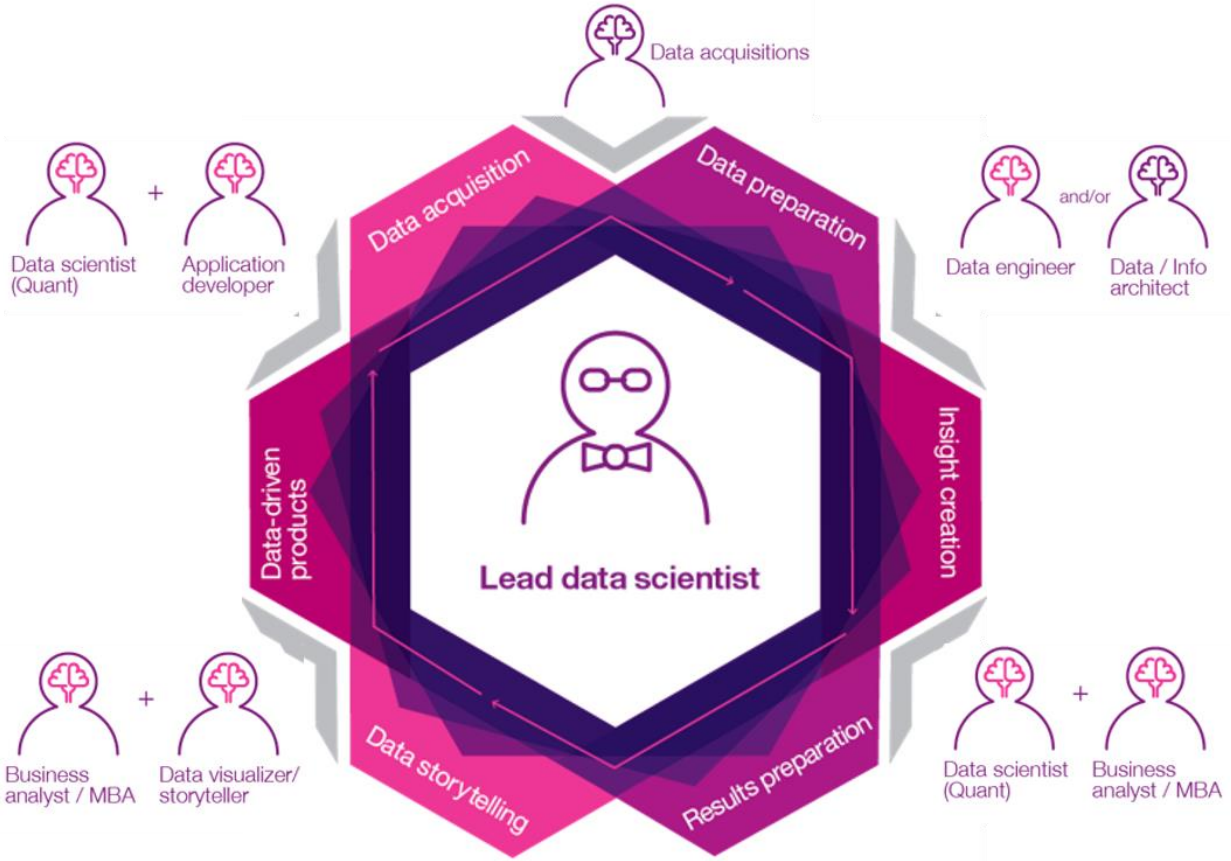
- (Managing) Data Science Architect (1)
- Data Scientist (1), Data Analyst (1)
- Data Science Application programmer (2)
- Data Infrastructure/facilities administrator/operator: storage, cloud, computing (1)
- Data stewards, curators, archivists (3-5)

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

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



EXAMPLE: IBM emerging professions



[ref] Mastering the art of data science: How to craft cohesive teams that create business results, IBM Institute for Business Value, 2016

- Create data science teams with varied backgrounds and skills

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

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)

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)

Further Developments and Actions

- Run surveys and key interviews for target communities
  - Primarily experts, visionaries and top executives at universities and companies
- Work with champion universities to practically validate the proposed EDSF
- Run a serial of the Champion Universities conferences: Spring 2017 and Summer 2017
  - Extend scope to Champions, Adopters, Ambassadors and Professionals
- Involve academic and industry experts and professional organisations to the definition of DS-BoK following from CF-DS
- Formally provide suggestions to ESCO for the definition of the Data Science professional profiles (occupations) family
- Formally provide suggestions for e-CF3.0 extensions for Data Science to CEN/PC 428
- Suggest required ACM CCS(2012) Classification extensions and proposal for Data Science curriculum definition

Discussion

- Questions
- Comments

- Invitation to contribution and cooperation:
  - Forum, EDISON Liaisons Groups, Champions Conference (Spring & Summer 2017)

- 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

- Survey Data Science Competences: Invitation to participate
  https://www.surveymonkey.com/r/EDISON_project_-_Defining_Data_science_profession

Definitions (according to e-CFv3.0)

- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results.
  - Competence vs Competency (e-CF vs ACM)
    - Competence is ability acquired by training or education (linked to learning outcome)
    - Competency is similar to skills or experience (acquired feature of a person)
- Competence is not to be confused with process or technology concepts such as, ‘Cloud Computing’ or ‘Big Data’. These descriptions represent evolving technologies and in the context of the e-CF, they may be integrated as elements within knowledge and skill examples.

- **Knowledge** in the context of competence definition is treated as something to know, to be aware of, familiar with, and obtained as a part of education.

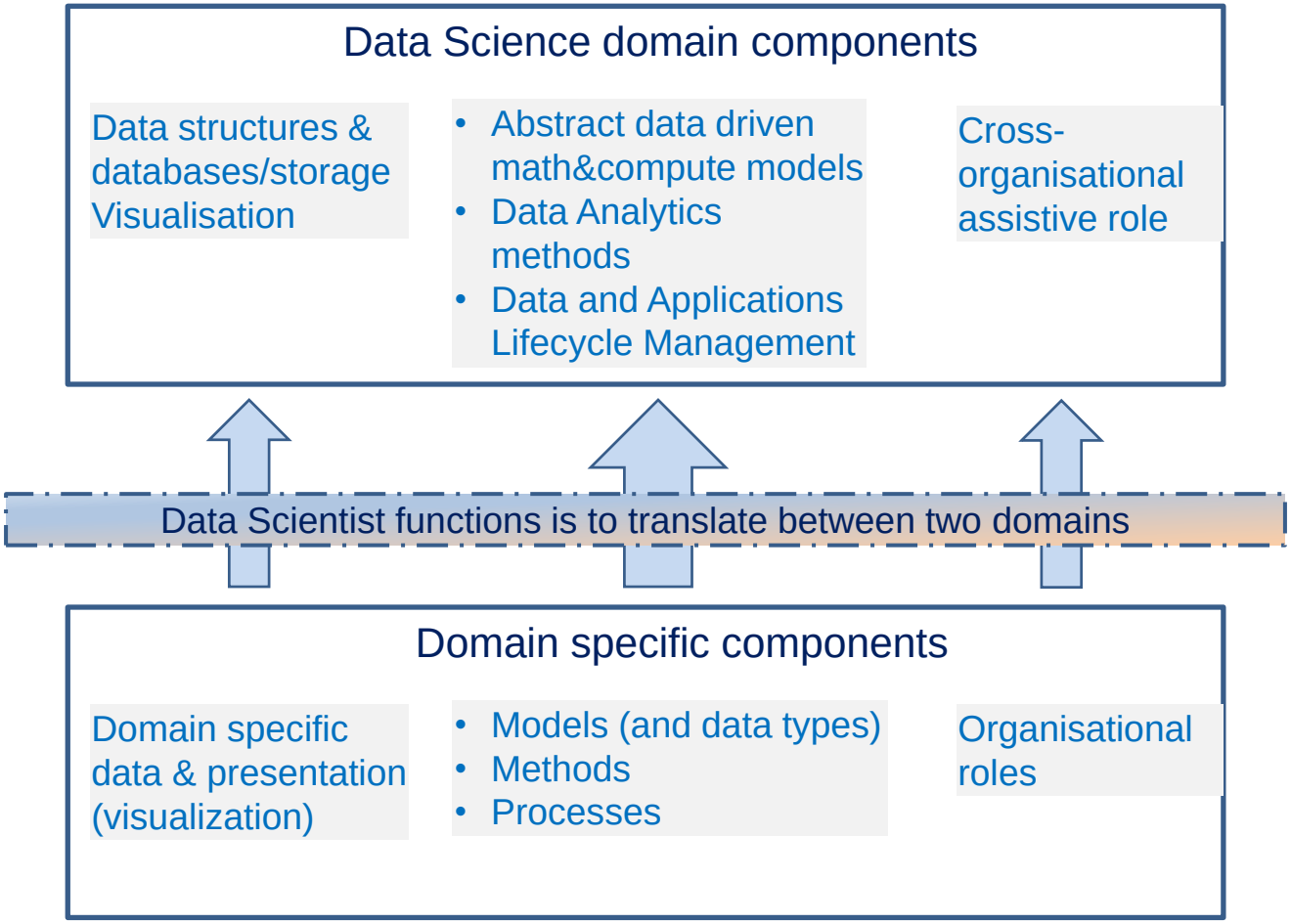
- **Skills** is treated as provable ability to do something and relies on the person’s experience.

Data Scientist and Subject Domain Specialist

- **Subject domain components**
  - Model (and data types)
  - Methods
  - Processes
  - Domain specific data and presentation/visualization methods
  - Organisational roles and relations

- **Data Scientist is an assistant to Subject Domain Specialists**
  - Translate subject domain Model, Methods, Processes into abstract data driven form
  - Implement computational models in software, build required infrastructure and tools
  - Do (computational) analytic work and present it in a form understandable to subject domain
  - Discover new relations originated from data analysis and advice subject domain specialist
  - Present/visualise information in domain related actionable way
  - Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data

Data Science and Subject Domains



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

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