

Training data scientists for industry

Adam Belloum



UNIVERSITEIT VAN AMSTERDAM



EDISON

building the data
science profession

Data Scientist: The Sexiest Job of the 21st Century

Thomas H. Davenport
Harvard Business Review

IBM 2013 report

Unleashing the potential of Big Data

- The process of incorporating Big Data into the operation of business, governance and education **will require hundreds of thousands of new, specially trained knowledge workers.**

McKinsey Quarterly (Feb 2016)

Big Data: getting a better read on performance

- About 40 percent of the profit improvements measured resulted from complementary and coordinated investments both in IT and in big data talent. **Skilled employees across the spectrum of data-analytics roles are in short supply**, so aggressive actions to address this problem are critical.

- ...

Windows of Opportunities

For educational institutions to start a curriculum in data science at all levels

**There is a real Need in
the job market of Data
scientist**

Rapidly growing
offers for training and
educating data
scientist

**Do we have a match between
the offer and the demand?**

Aim of the EDISON project

Coordination and Support and Action H2020 EU funded Project

**establish the data scientist as a profession
by *aligning industry needs with available
career paths***

- Support academies in developing curricula with respect to: expected profiles, required expertise and professional certification
- *Ensure research disciplines and market sectors coverage*
- *Gain consensus and engage with stakeholders*

EDISON actions and impact

IMPACT

Dramatically increase the number of data scientists

ACTIONS

Create a Data Science profession

Show opportunities for career path building

Interact with demand and supply sides

Competence Framework for Digital Single Market

Services to education and training

Support for accreditation and certification

Service for collaborating and sharing expertise and materials

Design model curricula

Engage stakeholder communities

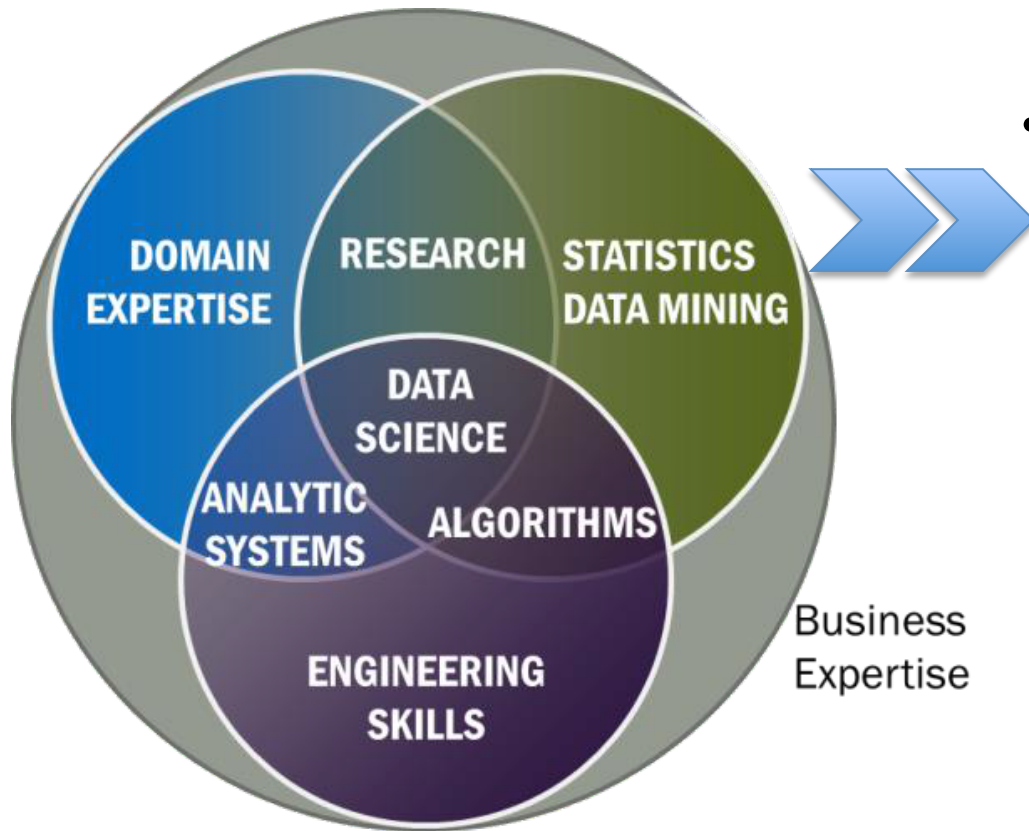
Sustain platforms of communities of practice

Organise "champion" universities

Interact with Expert Liaison Groups

EDSION Data Science Framework

Data Scientist mix of competences



- **Competence groups**

- Statistics and Data mining
- Engineering skills (computer related skills)
- Business expertise
- Domain expertise

Definition by NIST Big Data WG (2014-2015)

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**.

Provider / Consumer

- **Provider side (Program/training owners)**
 - *Accreditation of the programs*
 - *Increase the number of registration*
 - *Hire the appropriate experts*
- **Consumer side (learners, and HR)**
 - *Chose the program that get the first job*
 - *Chose the program that speedup carrier development*
 - *Hire the data scientist that fit exactly with needed profile*

Edison inventory for DS programs/courses/trainings



EDISON is a 2-year project (started September 2015) with the purpose of accelerating the creation of the Data Science profession.

Home Events News Library EDISON Contact EDISON Data Science Framework

Home | University Programs list

University Programs list

Country Language

Title

Apply

	Country	University	Language
Data Science	Spain	Barcelona Graduate School of Economics	English
track within Computer Science: Data Science and Technology track	Netherlands	Delft University of Technology	English
Data Science (new since sep 2014)	UK	Goldsmiths University of London	English
Data Science	UK	Heriot Watt University	English
Cross Disciplinary Studies Minor in Data Science	USA	California Polytechnic State University	English
Advanced Computing	UK	ImperialCollege London	English
Biomedical Research - Data Science track	UK	ImperialCollege London	English
Data Analytics	Canada	Western University Canada	
Predictive Analytics (E-learning)	USA	Northwestern University	English
Business Intelligence and Analytics	USA	Stevens Institute of Technology	English

Please get in touch to suggest new data science programs or alterations to the current list: s.brewer@soton.ac.uk

Latest news

Building the data science profession: workshop at DI4R 2016

Accreditation and certification schemes RDA 8th Plenary BoF meeting

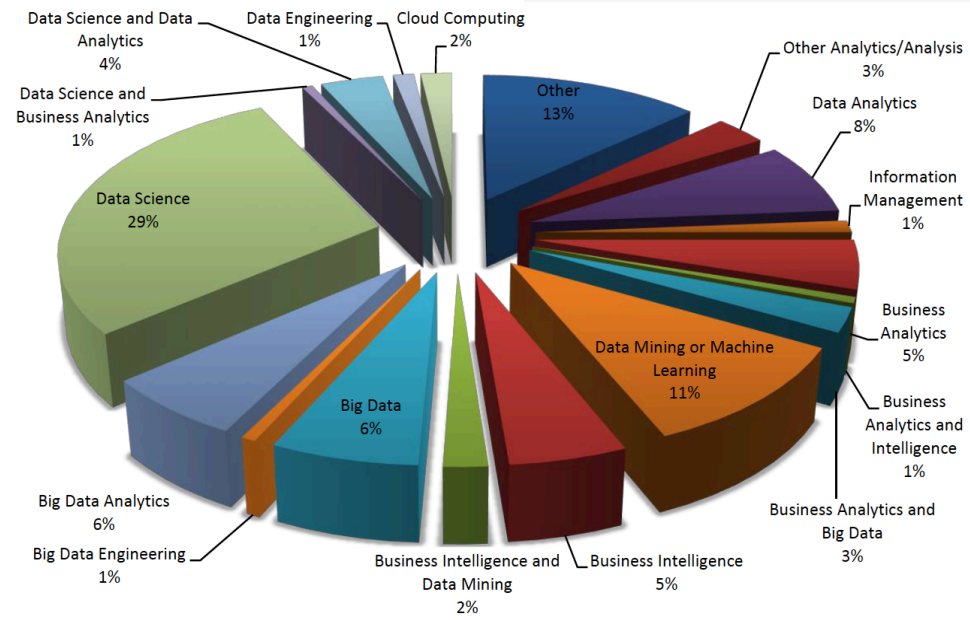
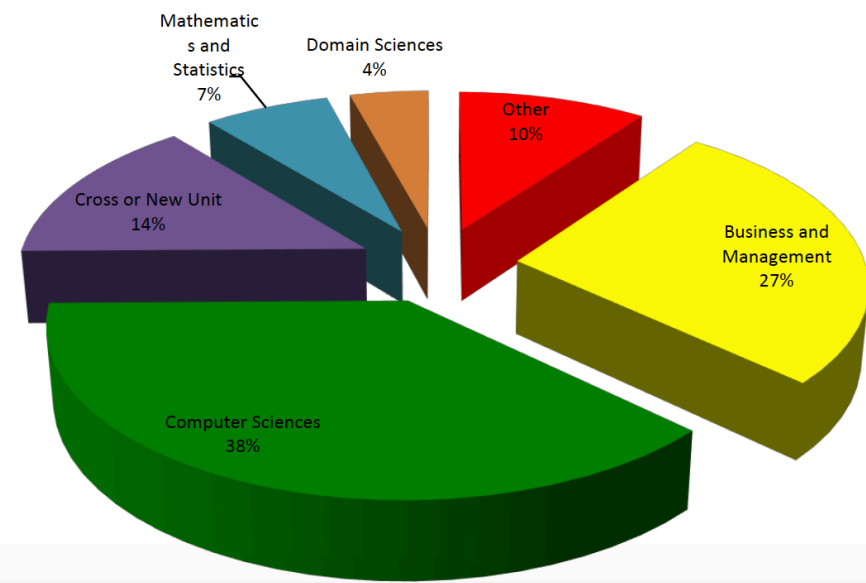
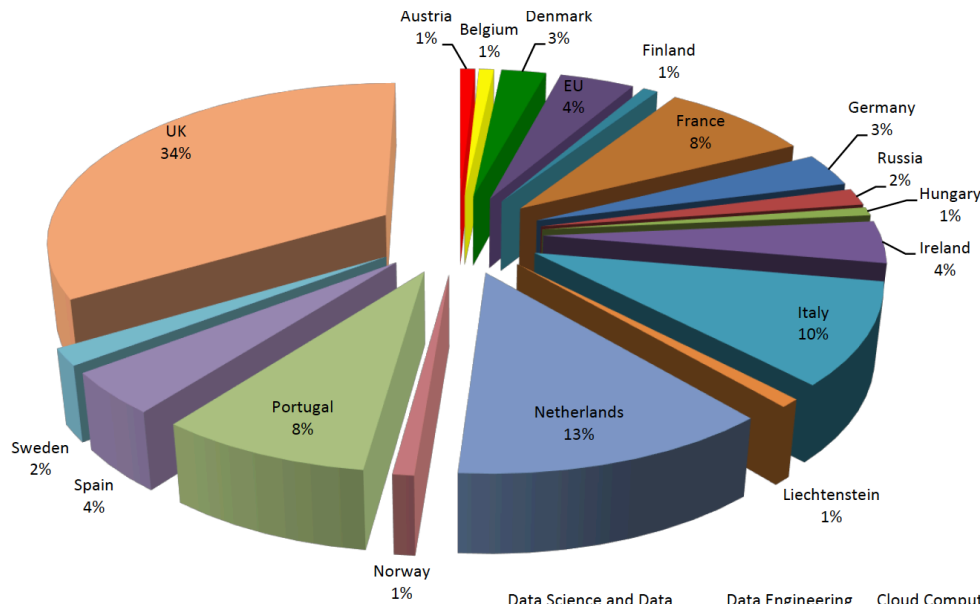
Second Education and Training Champions Conference: Madrid

EC launches New Skills Agenda for Europe

Engineering promotes the Master in Data Science at the University of Perugia

<http://edison-project.eu/university-programs-list>

Data Science Programs in EU

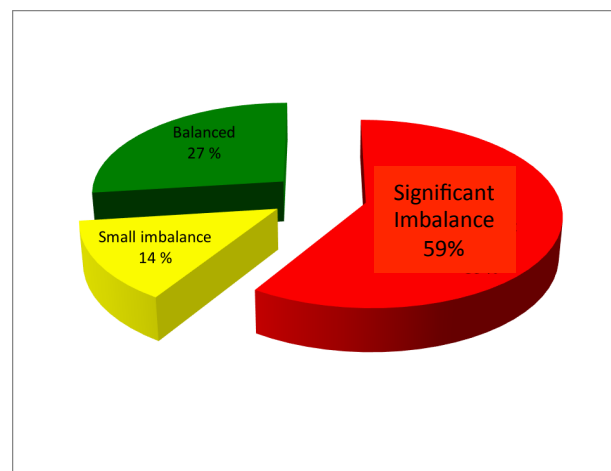
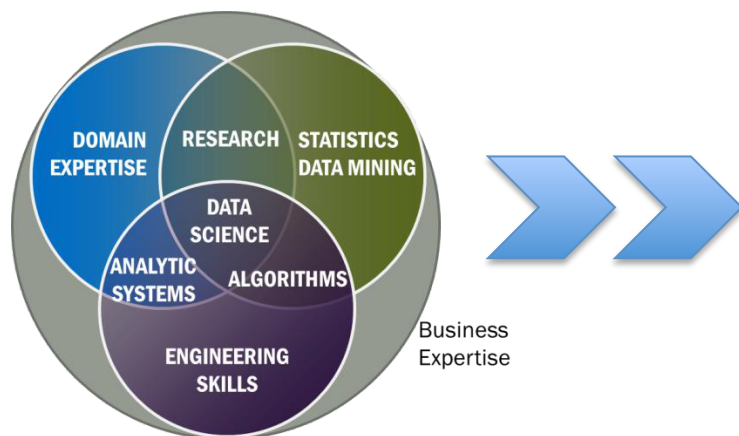


Program owner

Program Name

DS competence groups in existing DS program

- **Covering the DS competence groups**
 - *27% of the EU DS programs cover the 3 DS competences groups*

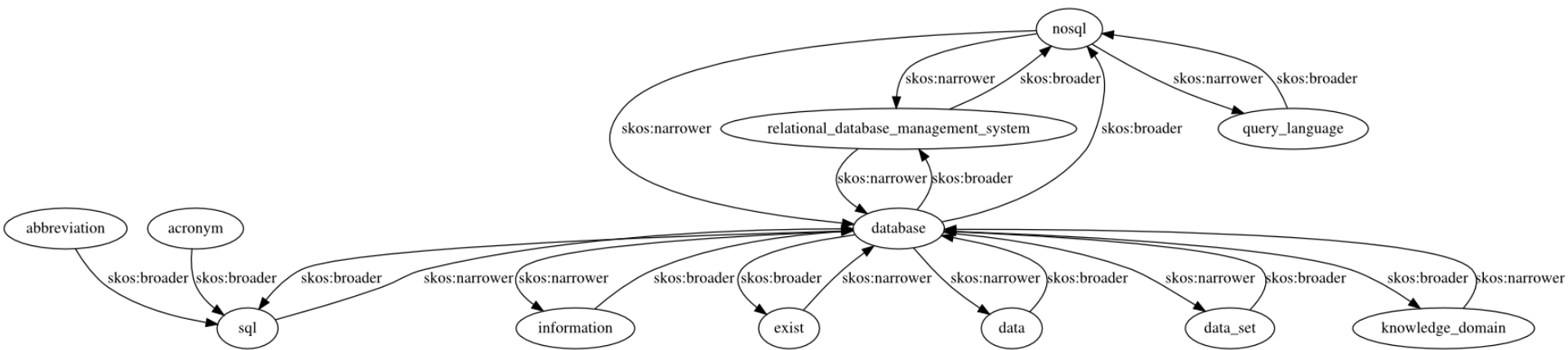


Notes:

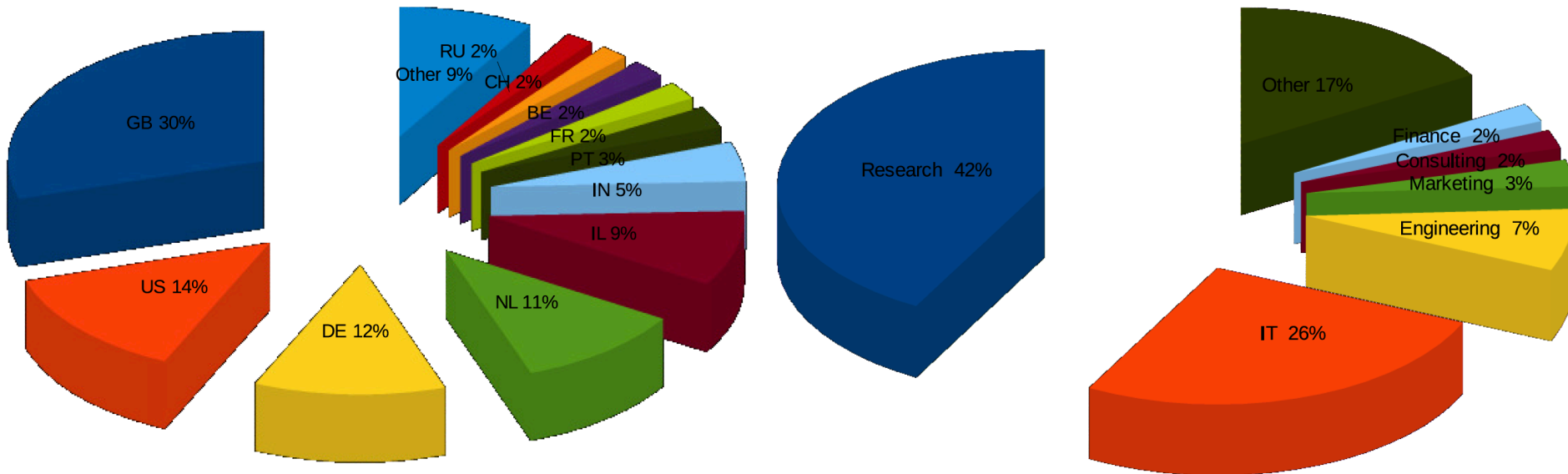
- *Most of the balanced programs are owned by multiple units*
- *35% of the Non- EU DS programs cover the 3 DS competence groups*

DS competence groups in existing DS job advertising

- term frequency count
 - too noisy, require a lot manual cleansing
- hierarchical relation discovery using hypernym-hyponym



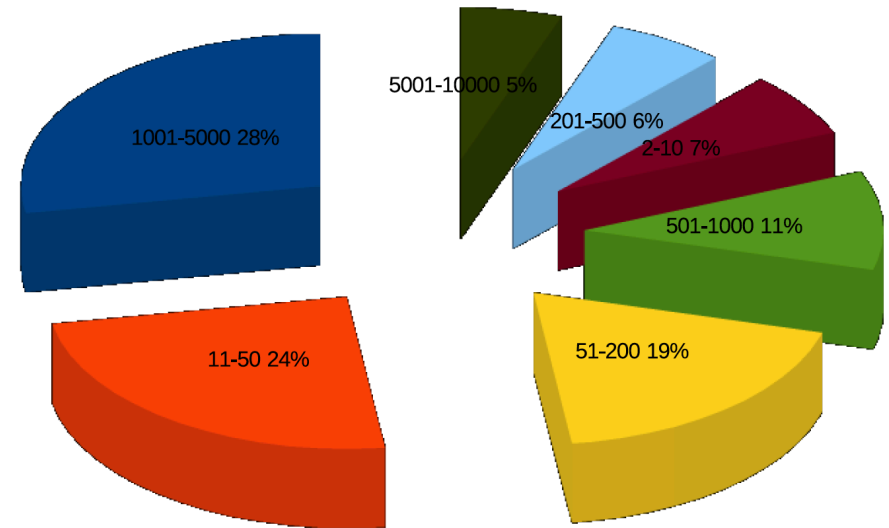
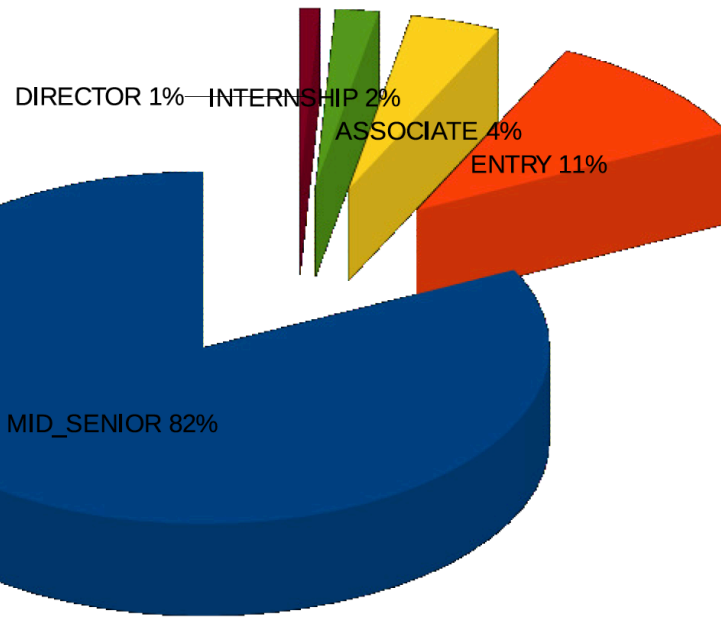
DS Job Market: analyzing of Data Science Job Ads



(a) Employer's country locations.

(b) Job position's function.

DS Job Market: analyzing of Data Science Job Ads

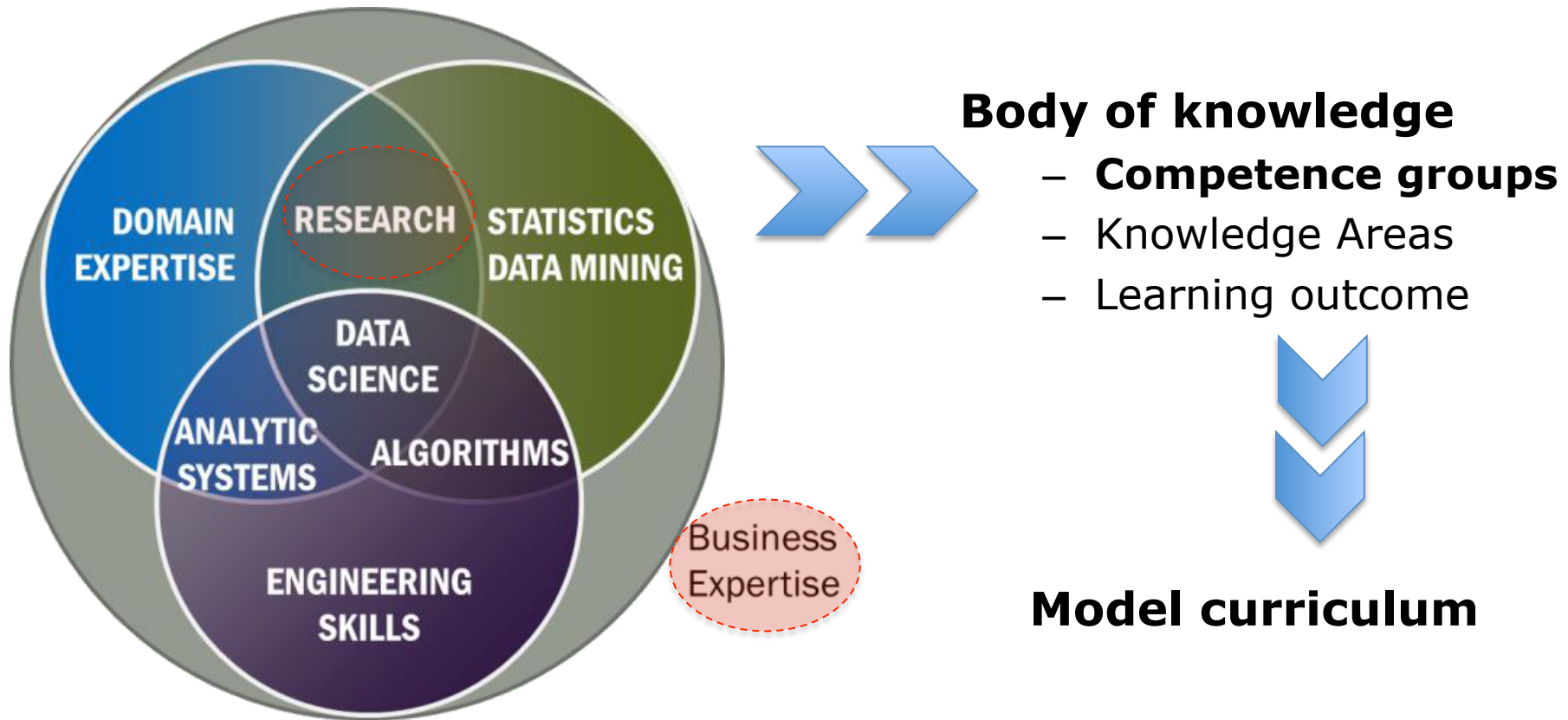


(c) Experience level required. (d) Employer's size in number of employees.

10 Top terms extracted with a frequency term and a hybrid method

Rank	Term Extraction	Hybrid Term Extraction
1	data	data_scientist
2	experience	communication_skills
3	skill	data_sets
4	model	data_analysis
5	scientist	data_science
6	learn	data_sources
7	big_data	data_analytics
8	science	data_analyst
9	customer	ideal_candidate
10	product	computer_science

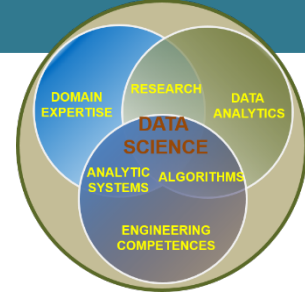
Data Scientist mix of competences



Definition by NIST Big Data WG (2014-2015)

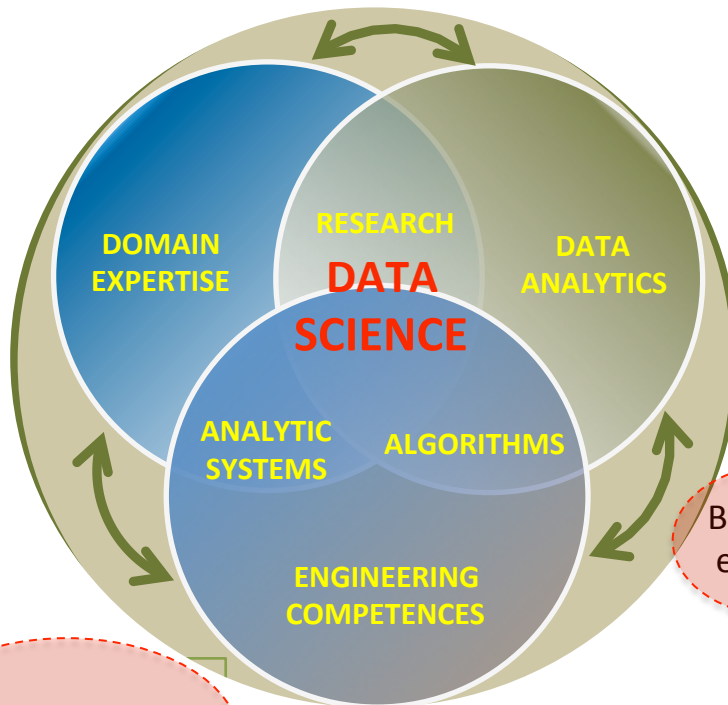
A **Data Scientist** is a practitioner who has *sufficient* knowledge in the overlapping regimes of expertise in **business** needs, **domain knowledge**, **analytical skills**, and **programming and systems engineering** expertise to manage the end-to-end scientific method process through each stage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

Data Science Competences Groups – PM



Data Science Competence includes 5 areas/groups

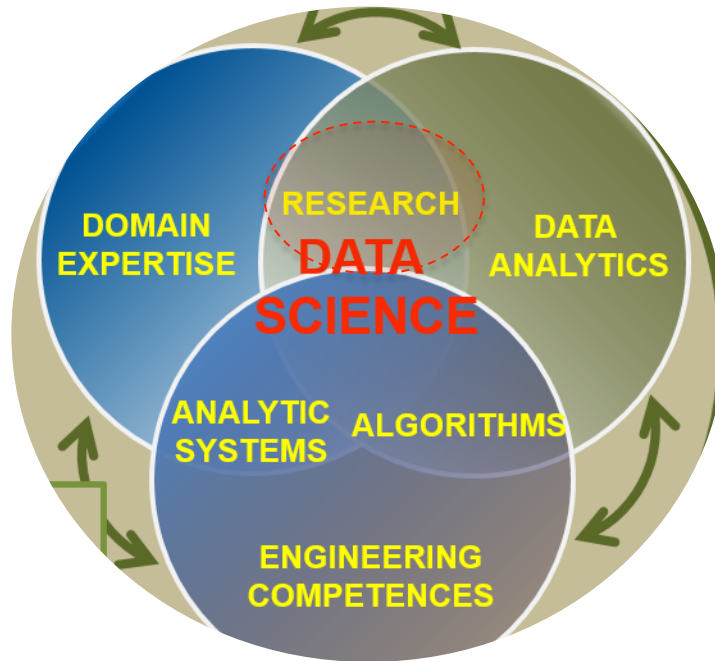
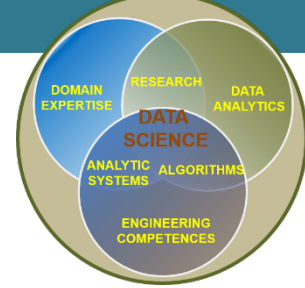
- Data Analytics
- Data Science Engineering
- Domain Expertise
- **Data Management**



Business Process Operations/Stages

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

Data Science Competence Groups - DM



Data Science Competence includes 5 areas/groups

- Data Analytics
- Data Science Engineering
- Domain Expertise
- **Data Management**

Business Process Operations/Stages

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

Scientific Methods

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

Business Process Management (for biz competences)

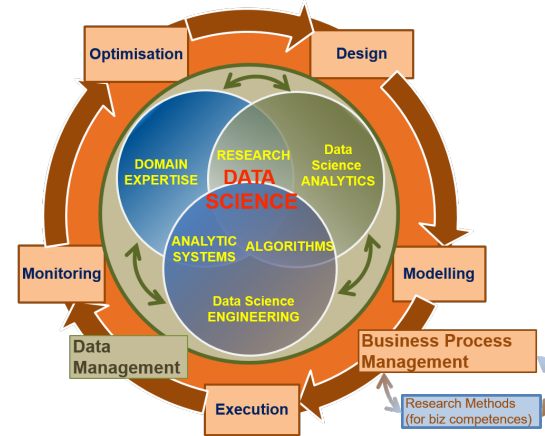
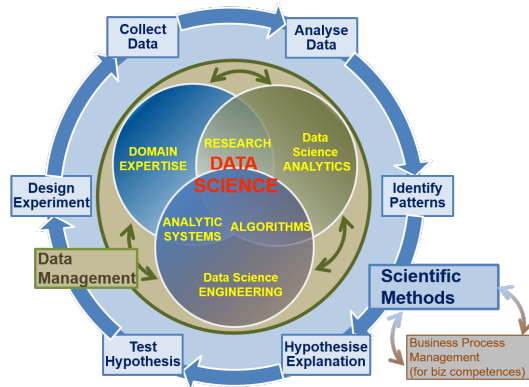
Data science Competence Groups

ESCO occupations family
Managers, Professionals, Technicians and associate
Professionals

PROFILE		DATA SCIENCE COMPETENCES GROUPS					Data Science Profile Definition [Table 5 in Section 4.2, D2.2]	
group	ID	Profile title	DSDA	DSDM	DSENG	DSRM	DSDK	
	DSP01	Data Science (group) Manager	3	4	3	3	2	Proposes, plans and manages functional and technical evolutions of the data science operations within the relevant domain (technical, research, business).
	DSP02	Data Science Infrastructure Manager	2	4	4	2	2	Proposes plans and manages functional and technical evolutions of the big data infrastructure within the relevant domain (technical, research, business).
	DSP03	Research Infrastructure Manager	2	4	4	3	2	Proposes plans and manages functional and technical evolutions of the research infrastructure within the relevant scientific domain.
	DSP04	Data Scientist	5	3	4	5	3	Data scientists find and interpret rich data sources, manage large amounts of data, merge data sources, ensure consistency of data-sets, and create visualisations to aid in understanding data. Build mathematical models, present and communicate data insights and findings to specialists and scientists, and recommend ways to apply the data.
	DSP05	Data Science Researcher	4	3	2	5	4	Data Science Researcher applies scientific discovery research/process, including hypothesis and hypothesis testing, to obtain actionable knowledge related to scientific problem, business process, or reveal hidden relations between multiple processes.
	DSP06	Data Science Architect	4	3	5	3	3	Designs and maintains the architecture of Data Science applications and facilities. Creates relevant data models and processes workflows.
	DSP07	Data Science (Application) Programmer/Engineer	4	2	5	3	4	Designs/develops/codes large data (science) analytics applications to support scientific or enterprise/business processes.
	DSP08	Data Analyst	5	3	3	3	4	Analyses large variety of data to extract information about system, service or organisation performance and present them in usable/actionable form
	DSP09	Business Analyst	5	3	3	4	5	Analyses large variety of data Information System for improving business performance.
	DSP10	Data Stewards	3	5	3	3	3	Plans, implements and manages (research) data input, storage, search, presentation; creates data model for domain specific data; support and advice domain scientists/ researchers
	DSP11	Digital data curator	1	5	2	2	3	Finds, selects, organises, shares (exhibits) digital data collections, maintains their integrity, up-to-date status and freshness, discoverability
	DSP12	Digital Librarians	2	5	2	2	3	Selection, acquisition, organization, accessibility and preservation of digital information. Manages digital materials, takes a lead role in the creation, maintenance and stewardship of digital collections, including the digitization of special collections. Develops strategies for effective management and preservation of library digital assets.
	DSP13	Data Archivists	1	5	1	1	3	Maintain historically significant collections of datasets, documents and records, other electronic data, and seek out new items for archiving.
	DSP14	Large scale (cloud) database designer	2	4	4	3	3	Designs/develops/codes large scale data bases and their use in domain/subject specific applications according to the customer needs.
	DSP15	Large scale (cloud) database admin	2	4	3	2	3	Designs and implements, or monitors and maintains large scale cloud databases
	DSP16	Scientific database administrator	2	4	3	2	3	Designs and implements, or monitors and maintains large scale scientific databases
	DSP17	Big Data facilities Operator	1	4	4	2	3	Manages daily operation of facilities, resources, and responds to customer requests. Includes all operations related to data management and data lifecycle
	DSP18	Large scale (cloud) data storage operator	1	4	3	1	1	Manages daily operation of cloud storage, Including related to data lifecycle, and responds to requests from storage users
	DSP19	Scientific database operator	1	4	3	2	3	Manages daily operation of scientific databases, Including related to data lifecycle, and responds to requests from database users

ESCO = European skills, competences, qualification and occupation

Definitions Body of knowledge / Knowledge Areas / Learning outcomes



Data Science KA (Knowledge Area groups)	alignment with existing BoK	Organization which developed the existing BoK
Data Science Analytics (DSA)	Business Analytics-BoK	Intentional Institute of Business Analysis http://www.iiba.org/babok-guide.aspx
Data Science Engineering (DSE)	Software-Engineering-BoK ICT professional-BoK,	http://www.ecompetences.eu/cen-ict-skills-workshop/ IEEE computer Society, SO/IEC TR 19759:2005 CEN ICT Skills Workshop
Data Management (DM)	Data Management-BoK	Global Data Management Community https://www.dama.org/content/body-knowledge
Scientific an Research methods (DSRM)	ACM-Computer science – BoK	ACM Association for Computing Machinery https://www.acm.org/education/CS2013-final-report.pdf
Business Process management (DSBP)	Project Management-BoK	Project management Institute
Data Science Domain Knowledge (DSDK)	--	

EDISON Data Science Framework (DSF)

- Data Science competence framework (DS-CF)
 - <http://edison-project.eu/data-science-competence-framework-cf-ds>
- Data science body of knowledge (DS-BoK)
 - <http://edison-project.eu/data-science-body-knowledge-ds-bok>
- Data Science Model Curriculum (DS-MC)
 - <http://edison-project.eu/data-science-model-curriculum-mc-ds>
- Data Science Professional profiles (DSP profiles)
 - <http://edison-project.eu/data-science-professional-profiles-definition-dsp>

EDISON Data Science Framework (DSF)



EDISON Discussion Document

Data Science Model Curriculum (MC-DS): Approach and Working version

Project acronym: EDISON
Project full title: Education for Data Intensive Science to Open New science frontiers
Grant agreement no.: 675419

Date	9 September 2016
Document Author/s	Yuri Demchenko, Adam Belloum, Tomasz Wiktorski
Version	0.2
Dissemination level	PU
Status	Working document, request for comments
Document approved by	



This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

<http://edison-project.eu/data-science-competence-framework-cf-ds>



EDISON Discussion Document

Data Science Body of Knowledge (DS-Bok): Approach and Working version

Project acronym: EDISON
Project full title: Education for Data Intensive Science to Open New science frontiers
Grant agreement no.: 675419

Date	9 September 2016
Document Author/s	Yuri Demchenko, Andrea Manieri, Adam Belloum
Version	0.3
Dissemination level	PU
Status	Working document, request for comments
Document approved by	



This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

<http://edison-project.eu/data-science-body-knowledge-ds-bok>



EDISON Discussion Document

Data Science Competence Framework (CF-DS): Approach and Working Version

Project acronym: EDISON
Project full title: Education for Data Intensive Science to Open New science frontiers
Grant agreement no.: 675419

Due Date	
Actual Date	4 July 2016
Document Author/s	Yuri Demchenko, Adam Belloum, Tomasz Wiktorski
Version	0.7
Dissemination level	PU
Status	Working document, request for comments
Document approved by	



This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

<http://edison-project.eu/data-science-model-curriculum-mc-ds>

Application to existing DS-Cu

• Master

– Master track **Big Data Engineering**

- Focus on DSENG → **target profiles** → **DSP02-03, DSP04, DSP06-7,**

– Master track **Artificial Intelligence and Data Science**

- Focus on DSDA using AI technique → **target profiles** → **DSP04-09,**

– Master **Data Science**

- Focus on DSDA → **target DS-profile** → **DSP04-09**

– Master **Business Analytics**

- Focus on DSDA → **target DS-profile** → **DSP09**

– Master track **Big Data Business Analytics** (Econometrics)

- Focus on DSBM → **target DS-profiles** → **DSP09**

• Postgraduate

– Course/Training HPC and Big data

- Focus on DSENG → **target DS-profiles** → **DSP02-03, DSP04, DSP06-7, DSP14, DSP17,**

– 1 month (8 hours per week)

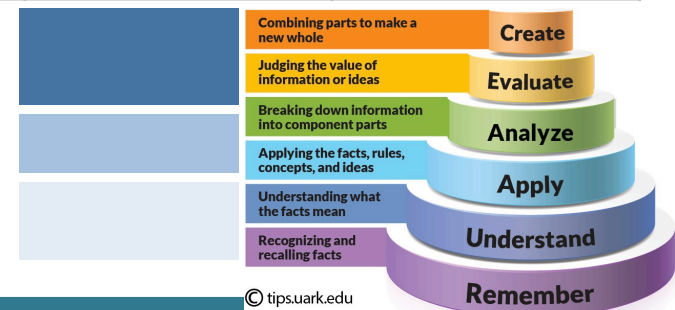
Track = Specialization

ESCO = European skills, competences, qualification and occupation

DSP01	Data Science (group) Manager
DSP02	Data Science Infrastructure Manager
DSP03	Research Infrastructure Manager
DSP04	Data Scientist
DSP05	Data Science Researcher
DSP06	Data Science Architect
DSP07	Data Science (Application) Programmer/Engineer
DSP08	Data Analyst
DSP09	Business Analyst
DSP10	Data Stewards
DSP11	Digital data curator
DSP12	Digital Librarians
DSP13	Data Archivists
DSP14	Large scale (cloud) database designer
DSP15	Large scale (cloud) database admin
DSP16	Scientific database administrator
DSP17	Big Data facilities Operator
DSP18	Large scale (cloud) data storage operator
DSP19	Scientific database operator

Which level of Knowledge is needed for each KA for each DS-profile

	Managers, Professionals, Technicians and associate Professionals				
	DSP01-DS03	DSP04-DS09	DSP10-13	DSP14-DS16	DSP17-DS19
Data analytics (DSDA)					
Data Science Engineering (DSENG)					
Data Management (DSDM)					
Scientific research & method (DSRM)					
Business process (DSBP)					
Domain Knowledge (DSDK)					

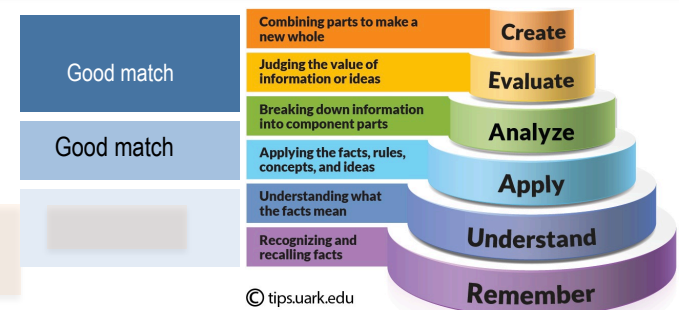


Application to existing DS-Curricula

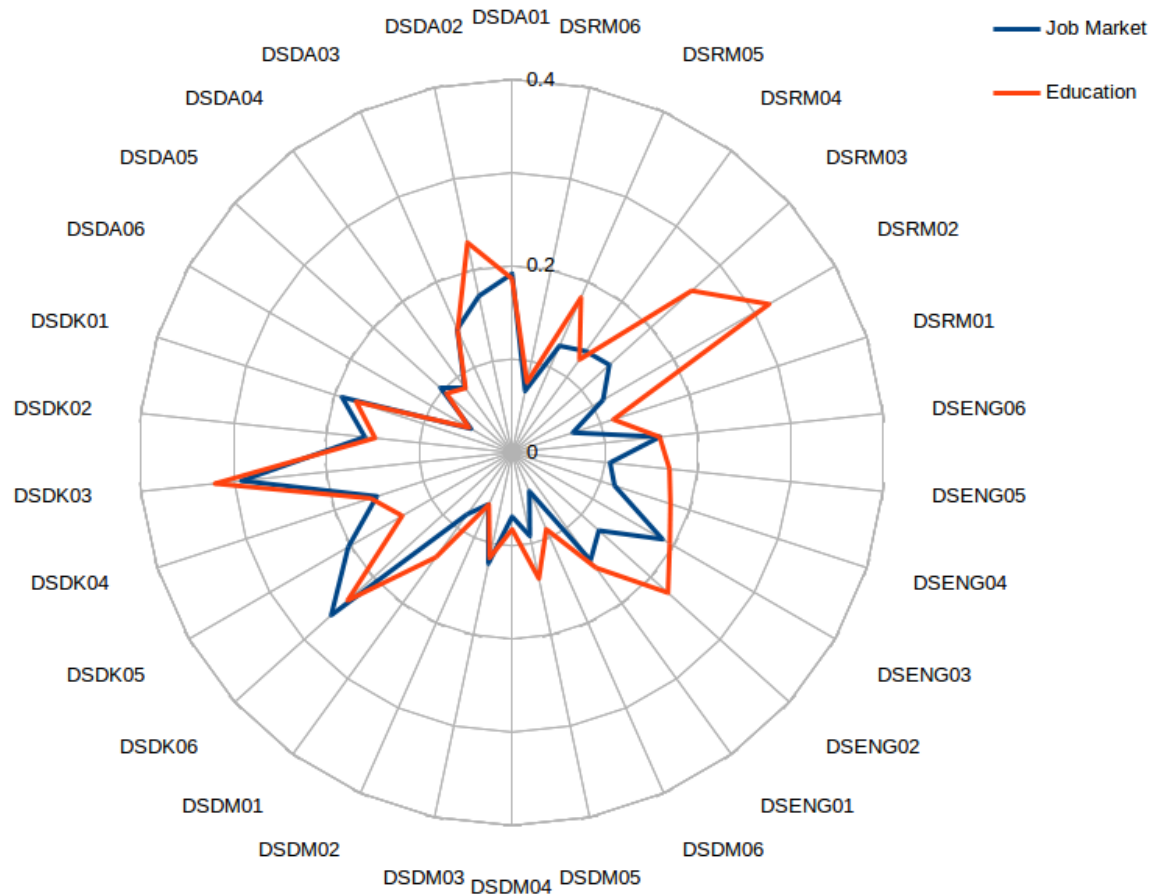
	Managers : DSP01-DS03	Professionals: DSP04-DS09	Professionals (data handling/management: DSP10-13	Professionals (database): DSP14-DS16	Technician and associate profession: DSP17-DS19
Data analytics	1,2,3 1,2,3 1,2,3	1,2,3 1,2,3 1,2,3 1,2,3 1,2,3 4,5			
Data Science Engineering	1,2,3 1,2,3 1,2,3	1,2,3 1,2,3 1,2,3 1,2,3 1,2,3 1,2,3			
Data Management		1,2,3 1,2,3 1,2,3 1,2,3 1,2,3 1,2,3			
Scientific research and method	1,2,3 1,2,3 1,2,3	1,2,3 1,2,3 1,2,3 1,2,3 1,2,3 1,2,3			
Business process					
Domain Knowledge					

1. Master track **Big Data Engineering**
2. Master track **Artificial Intelligence** and **Data Science**
3. Master **Data Science**
4. Master **Business Analytics**
5. Master track **Big Data Business Analytics** (Econometrics)

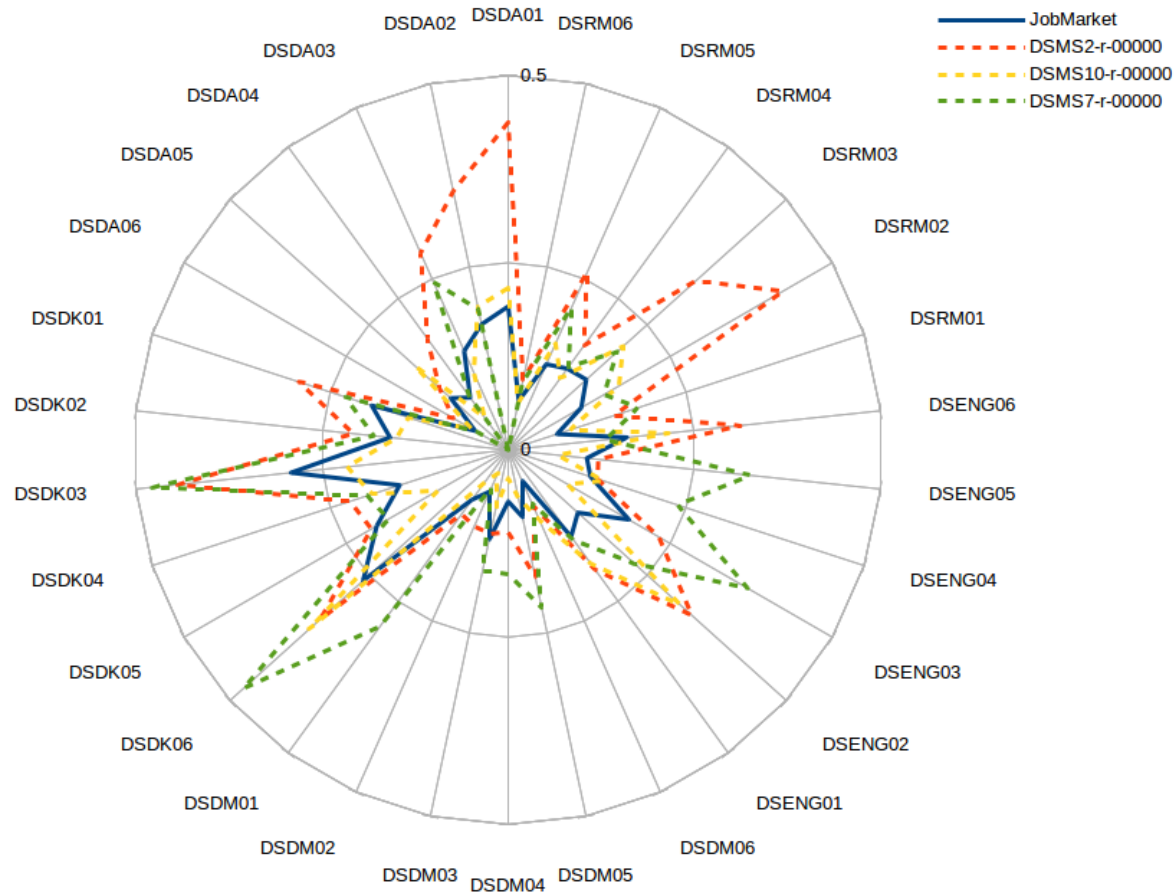
Target for short trainings and Course (which is modular and easily customizable)



Profile of job market and education for each of the 30 competence groups.



Profile of job market and individual courses for each of the 30 competence groups



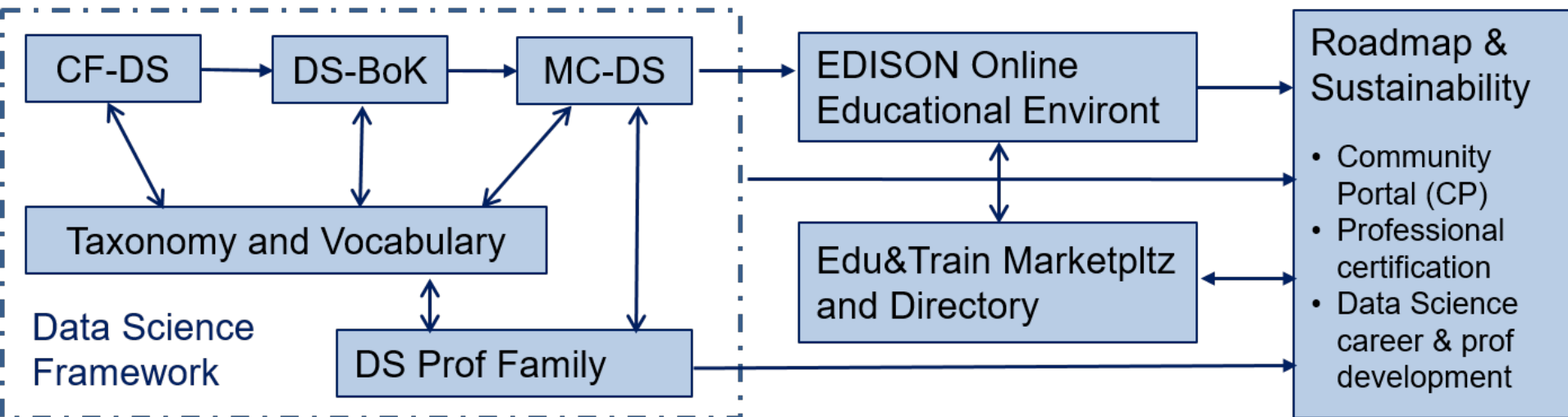
[Data Science Jobs descriptions collected from https://www.linkedin.com/jobs/](https://www.linkedin.com/jobs/)
[Data Science program collected from http://www.kdnuggets.com/education/index.html](http://www.kdnuggets.com/education/index.html)

What do we need ...

- **Data Science Curricula Foundation**

- *Competence Framework for Data Science (CF-DS)*
- *Data Science Body of Knowledge (DS-BoK)*
- *Model Curriculum for Data Science (MC-DS)*

*providers
(program developers)*








Competence Benchmark

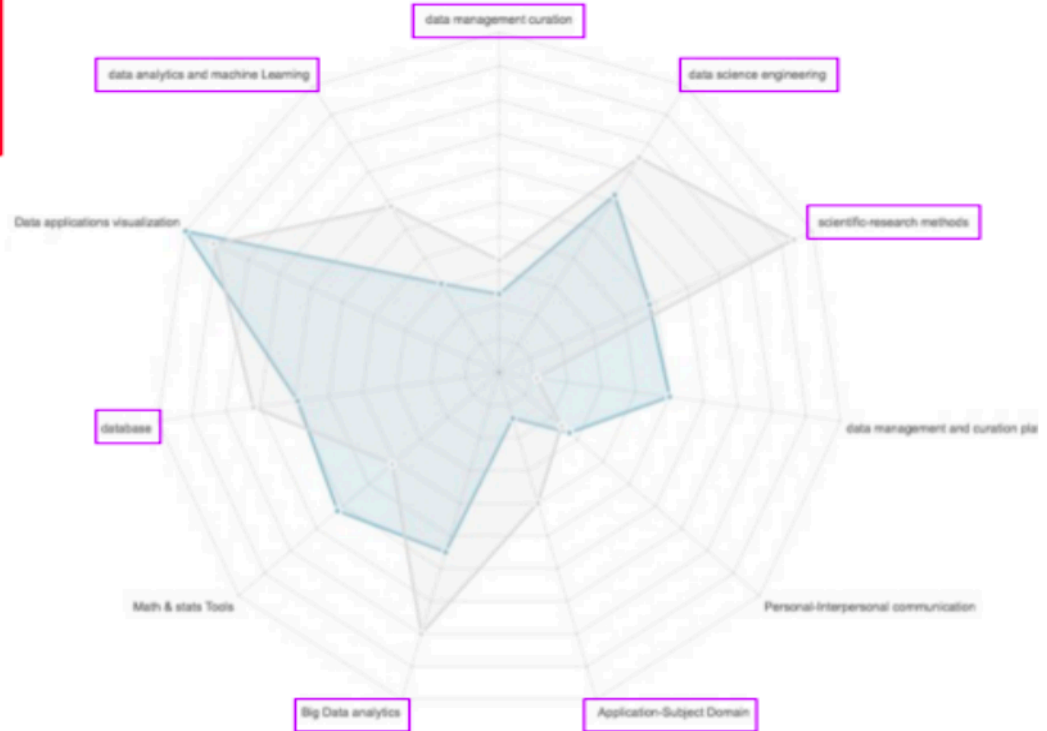


Your profile and the chosen profession "Data Analyst" match 83.8%. A tailored training programme is ready for you.

TRAINING PROGRAMME

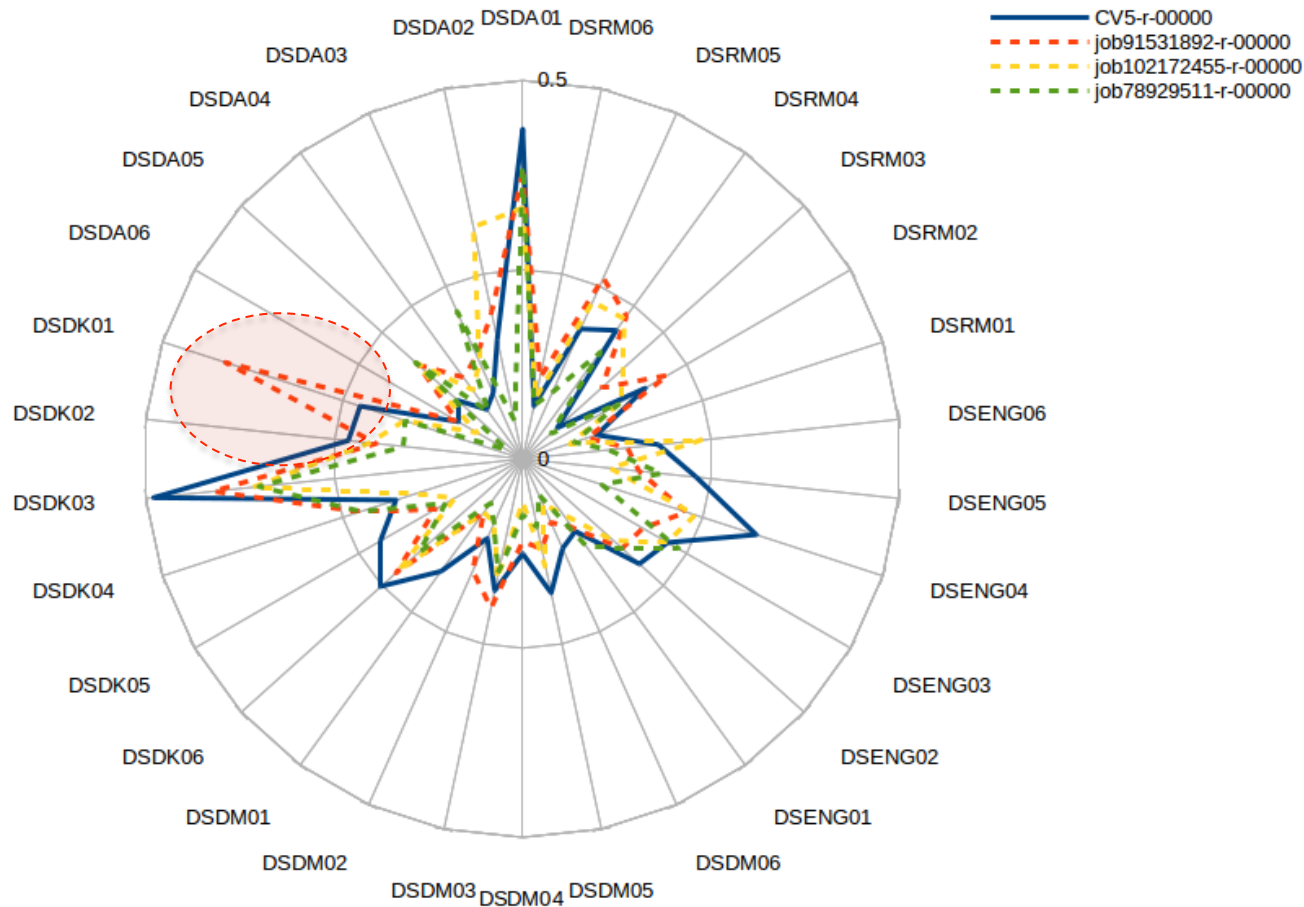
PROFILE GROUP	PROXIMITY
 Data Science Professionals Data Scientist Data Science Researcher Data Science Architect Data Science (Application) Programmer/Engineer Data Analyst Business Analyst	83.8% +
 Database professionals	61.8% +
 Data Handling / Management Professionals	40.5% +
 Technicians and associate professionals	35.5% +
 Data Science Infrastructure Manager	21.6% +

The depicted polygons are your declared competences, while the graphs are average values of each competence for a profile group. The gap between you and the chosen profile group is represented with two colors. Green areas indicate the competences that you have that exceed those required by the represented profile. Red areas indicate the competences required by the represented profile where you have a deficiency.



Competence Benchmark is an interactive web application that is able to assess individual competence profile and propose a tailored Data Science training programme.

Profile of CV and individual job profiles for each of the 30 competence groups



How can I engage with EDISON?

<http://edison-project.eu/>



Engagement and Interaction

The EDISON initiative has a number of channels for engagement depending on your needs in the Data Science profession. The EDISON Data Science Framework (EDSF) offers opportunities and benefits for managers, trainers, teachers, researchers, employers and Data Science professionals.

Through deeper understanding and greater familiarity with Data Science and the associated competences inherent in the profession, all stakeholders stand to gain something. Furthermore the EDISON initiative welcomes feedback and interaction in order to continue developing these resources that all stakeholders can enjoy.

How can I engage with EDISON?

Here are some of the current channels for interaction and engagement (although we welcome other approaches):

- Research Data Alliance (RDA) - <https://rd-alliance.org/>
 - Interest Group - **Education and Training on handling of research data (IG-ETRD)** - Chair(s): Yuri Demchenko, Laura Molloy, Amy Nurnberger, Christopher Jung
 - Birds of a Feather - **Accreditation and certification schemes**: first meeting will be at RDA 9 in Denver, Colorado, USA 17th September, 2016
 - Group chair serving as contact person Steve Brewer
 - Birds of a Feather - **Research Data Management Literacy**: first meeting will be at RDA 9 in Denver, Colorado, USA 16th September, 2016
 - Group chair serving as contact person Yuri Demchenko
- We are also collaborating with the following projects and organisations:
 - CODATA - <http://www.codata.org>
 - European Data Science Academy (EDSA) - <https://edsa-project.eu>
 - Technical and Human Infrastructure for Open Research (THOR) - THOR
 - DataLab (Scotland) - <http://www.thedatahub.com>

Social Media

Follow the EDISON project on Twitter: @EdisonEU - do share your posts with the @EdisonEU community where relevant.

The most recent Tweets can be seen on the EDISON website (use #datascience for relevant posts)

LinkedIn: join the EDISON group for updates and discussions: <https://www.linkedin.com/groups/8473188>

EDISON initiative

EDISON

- Data scientist profession
- EDISON Project
- Expert Liaison Groups - ELG
- Education and Training Champions
- National Action Plans
- **Engagement and Interaction**
 - Engagement coordination
 - Contact

Latest news

Building the data science profession: workshop at DI4R 2016

Accreditation and certification schemes RDA 8th Plenary BoF meeting

Second Education and Training Champions Conference: Madrid

EC launches New Skills Agenda for Europe

Engineering promotes the Master in Data Science at the University of Perugia

Tweets by @EdisonEU

EDISON EU project @EdisonEU
Yes, a good question, do we have to start with a definition of each? Or perhaps an understand of what each is not? twitter.com/evamern/status/...

EDISON EU project @EdisonEU
Impressive graphic from Kay Raseroka (@Rasawana) showing the many #data links reaching out from #Africa #RDAPlenary

Embed View on Twitter

References

1. Spiros Koulouzis, **Adam S.Z. Belloum**, and Marian T. Bubak “Data Access Profiles of VPH Applications” VPH workshop, Virtual Physiological Human conference (VPH 2016) Amsterdam 26-28 September 2016.
2. R. Cushing, **A.S.Z Belloum**, and M.T. Bubak, Towards A Data Processing Plane: An Automata-based Distributed Dynamic Data Processing Model, Volume 59, June 2016, Pages 21–32, <http://dx.doi.org/10.1016/j.future.2015.11.016>
3. S. Koulouzis, **A.S.Z Belloum**, Z. Zhao, M. Zivkovic, M.T. Bubak, and C. de Laat, SDN-Aware Data Transfers for Scientific Applications, Volume 56, March 2016, Pages 64–76 FGCS journal. 2015, PrePrints, doi: [10.1016/j.future.2015.09.032](http://dx.doi.org/10.1016/j.future.2015.09.032)
4. M. Baranowski, **A.S.Z Belloum**, M Bubak, Cookery: a Framework for developing Cloud Applications, 13th Annual IEEE International Conference on High Performance and computing and simulation (HPCS 2015), July 22-24, Amsterdam, The Netherlands.
5. J. Serrat, T. Szeplieniec, **A.S.Z. Belloum**, J. Rubio-Loyola, O. Appleton, T. Schaaf, J. Kocot, gSLM: The Initial Steps for the Specification of a Service Management Standard for Federated e-Infrastructures, 8th IFIP International Conference on Research and Practical Issues of Enterprise Information Systems CONFENIS 2014, Hanoi, Vietnam
6. R. Cushing, **A.S.Z Belloum**, M.T Bubak, A. Oprescu, C.T.M. de Laat, Exploratory Data Processing Using Non-deterministic Finite Automata, workshop on Large Scale Distributed Virtual Environments on Clouds and P2P - LSDVE 2014 held in conjunction of Euro-Par 2014 Porto, Portugal.
7. S. Koulouzis, D. Vasyunin, **A.S.Z Belloum**, and M.T. Bubak, Data Storage Federation for VPH Applications, submitted to Virtual Physiological Human Conference 2014, T rondheim September 9-12, 2014
8. Artem Chirkin, **A.S.Z. Belloum** and S. V. Kavalchuk, Execution Time Estimation fro workflow scheduling, workshop on Workflows in support for large scale Science, Held in conjunction with SC 14 and in-cooperation with SIGHPC, Nov. 16, 2014, New Orleans, Louisiana.