The 2019 High Performance Computing & Simulation International Conference
(HPCS 2019)
Dublin, Ireland
July 15 - 19, 2019

TUTORIAL

Becoming a Data Scientist and Educating Data Scientists:

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

Yuri Demchenko
University of Amsterdam

TUTORIAL DESCRIPTION

Data Science is an emerging field of science, which requires a multi-disciplinary approach and has a strong link to Big Data and data driven technologies that created transformational effect to all research and industry domains. There is a critical gap in current supply of Data Scientists and other Data Science and Analytics (DSA) enabled professions in research, industry and government. Thousands of Data Scientist and related vacancies remain unfilled for months. Companies are looking to specialists who will help them to make the company data driven and benefit from the new technologies from Big Data, supercomputing, IoT to machine learning and cognitive technologies. However, getting the Data Science position is not easy and requires extensive knowledge and experience in many areas that comprise modern Data Science.

The education and training of Data Scientists currently lacks a commonly accepted, harmonized instructional model that reflects by design the whole lifecycle of data handling in modern, data driven research and the digital economy.

To address this problem, the tutorial will start from the definition of the Data Scientist that is based on the extended NIST SP1500-1 definition: “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.”

The competences required from the Data Scientists to successfully work in different work environments in industry and in research and through the whole career path include:

- Data Analytics including statistical methods, Machine Learning and Business Analytics
- Data Science Engineering: software and infrastructure
- Data Management and Governance
- Research Methods and Project Management
- Subject Domain competences and knowledge

This tutorial introduces the EDISON Data Science Framework (EDSF) that provides a foundation for the Data Science profession definition. The EDSF includes the following core components: Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), and Data Science Professional profiles (DSP profiles). The MC-DS is built based on CF-DS and DS-BoK, where Learning Outcomes are defined based on CF-DS competences and Learning Units are mapped to Knowledge Units in DS-BoK. In its own turn, Learning Units are defined based on the ACM Classification of Computer Science (CCS2012) and reflect typical courses naming used by universities in their current programmes.

The EDSF also defines the Data Science professional skills and 21st Century skills that are generally required by modern data driven companies.

For educators, the tutorial provides examples how the proposed EDSF can be used for designing effective Data Science curricula as well as individual competences assessment and Data Science teams building.

For job seekers, the tutorial will advise how to understand a vacancy description, understand what the company actually needs and how to successfully manage job application and interview.

Tutorial OUTLINE

The tutorial will cover the following topics:

- What is Data Science and related technologies
- EDISON Data Science Framework (EDSF)
- Data Scientist and Data Science related competences and Skills
- Data Science professional skills and 21st Century skills also known as workplace or “soft” skills
- Data Science Body of Knowledge and Data Science Model Curriculum
- Example Data Science competences benchmarking and tailored Data Science curriculum design
- Digital skills and Industry 4.0, digital transformation of organisations
- How to develop your DSA related competences and skills
- How to apply for Data Science job? What competences and skills are essential?

REFERENCES

EDISON Data Science Framework (EDSF), EDISON Community Initiative [online]
https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome
https://github.com/EDISONcommunity/EDSF/

Yuri Demchenko, et al, EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry, 3rd IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2016).

Yuri Demchenko, Luca Comminiello, Gianluca Reali, Designing Customisable Data Science Curriculum Using Ontology for Data Science Competences and Body of Knowledge. Proceedings ICBDE2019 Conference, 31 March – 1 April 2019, Greenwich, London.

The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4 [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

REQUIREMENTS AND TARGET AUDIENCE

No special requirements to audience.
There is no specific knowledge of Data Science or Big Data required.
The expected target audience is wide but primarily oriented on two groups in the Big Data and Data Science ecosystem: educators and course developers; and practitioners who already work as Data Scientists and those who want to become as Data Scientist.
Manager and HR workers may benefit from the presented tools for job description generation and candidates' resume assessment methodologies and tools

TUTORIAL DURATION

The tutorial material will be presented in one 2 hours sessions.

A/V AND EQUIPEMNT

Standard presentation facilities, no AV required.

INSTRUCTOR BIOGRAPHY AND PHOTO

Yuri Demchenko is a Senior Researcher at the System and Network Engineering of the University of Amsterdam. He is graduated from the National Technical University of Ukraine "Kiev Polytechnic Institute" where he also received his PhD (Cand. of Science) degree. His main research areas include Data Science and Data Management, Big Data and Infrastructure and Technologies for Data Analytics, DevOps and cloud based software development, general security architectures and distributed access control infrastructure for cloud based services and data centric applications. He is currently involved in the European projects GEANT4, MATES, FAIRsFAIR where he develops different elements of cloud based infrastructures for scientific research, and issues related to Data Science and digital skills development. Yuri has coordinated the EU funded EDISON project (2015-2017) which has developed the EDISON Data Science Framework (EDSF) that provides a conceptual foundation and practical basis for building the Data Science profession. His recent research are also extending into data economics and open data market models.

He is actively contributing to the standardisation activity at RDA, OGF, IETF, NIST, CEN on defining Big Data Architecture Framework, Data Science competences, and data properties as economic goods.