Table of Contents

The EDISON project

Notes by Maria Dimou - CERN representative to the project.

Mandate

The Horizon 2020 EU-funded EDISON project aims at defining the new Data Scientist profession. For universities to define their courses. For companies to define the required competencies. By: defining the Data Scientist Profile, providing a web-based tool to (self-)evaluate one's compliance with the profile, providing a basis for a Data Science professional certification. It is a 2-year project (started September 2015).

Definitions

Who is a Data Scientist?

A practitioner who has sufficient knowledge to dig through the life cycle of Big Data till the delivery of scientific and business results of value for science and industry. He/she has a focused interest in data digging, while looking for something **usable**. Communication skills are very important in this profession, also the innovation potential, the cost reduction etc. Data scientists span from Data Analyst to system administrator and librarian. Check here existing NIST (ACM IEEE) work on the subject and more. The table of relevant current and future professions is so large that we should better talk about Data ScienceS and not Data Science (as we say Natural Sciences) A survey will be out soon to poll various specialists' communities for input on the Data Science professional. The table of specialisations will be debated with ESCO (European Skills, Competences, Qualifications and Occupations) framework and platform

Why is this new profession important?

Because data analysis, statistics and data mining will "discover hidden and obscure relationships between processes and events, which will lead to new discoveries and innovation".

Links

- EDISON document repository.
- EDISON survey on the Data Scientist's job definition.
- CERN Representative's reply to the survey.
- Forum on Data Analytics and Tomorrow's workforce **20170330 but also recorded** by The Atlantic.

EDISON Champions Conference - Warsaw 2017/06/19-20

Highlights relevant to CERN

- The Champions Conference brings together experts in Data Science definition from education, research and consulting.
- The EDISON project ends in August 2017. The Community portal will be maintained.
- The project officer Steve Brewer plans to continue by bringing together experts consulted so far.
- Maria Dimou proposes to invite him for a talk in the autumn, so that the benefits for continuous involvement by CERN can be understood.
- CERN has a lot of experience to contribute in the development of a toolkit identifying one's skills and training needs in the area of Data Science. The benefit for CERN will be in better defining the IT job openings.

Conclusions on the Data Scientist (DS) definition: The DS is a professional able to:

- Think about all possible uses of the data short & long term
- Make a defend a plan to handle business risks based on data analysis & projection
- Deal with uncertainty and make policy / take decisions.
- Know how to select tools to ensure data quality

Justification for the above: The technology ages, the awareness of data value, impact & ethics doesn't.

Recommendations for educational programmes also by employers:

- One often becomes a DS as a 2nd degree/specialisation, so:
- Ensure a common basic technical curriculum but also...
- Acknowledge & integrate the wealth of different backgrounds/experiences.
- Teach business practices in addition to the academic study plan.
- Make sure the data life-cycle (from production to archiving) is taught with emphasis.
- Cultivate communication, team work, meetups where learning comes out of success stories.
- Be convinced that learning is work and should be a continuous process.

Links:

1. Event agenda. Material will be uploaded early July 2017.
2. twitter #EDISONWarsaw for the event.
3. Updated material Poster & FAQ
4. All library documents, especially EDISON Data Science Framework
5. Interesting related activities - The Data Carpentry Project.
6. Progress since the Krakow meeting in September 2016:
    1. A PhD student with help by an IBM expert is conducting interviews with dutch ministries to evaluate the computing "literacy" of their elected members.
    2. Proposing to CERN to become another pilot tester of the **toolkit**.
    3. Updated the Framework document in the repository HERE (by Yuri Demchenko)
    4. CERN's answers to the Data Scientist Survey (by M.Dimou)

In more detail

- The conference host Professor Marek Niezgódka, Director of ICM, Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, reported on a successful collaboration by 20 polish organisations for the creation of *Digital Poland*, to define a national infrastructure for research. There is a strong cohesion between academia and industry, via a national e-campus 10Gbps Geant network backbone and 5 HPC centres with 1-2 PF capacity each. This is an important and urgent effort as the rapid growth of the polish economy exceeds *awareness adjustment* of public administration and industry. Several studies' curricula contain a DS profile e.g. the one on 'Computational Engineering'. Advice to public administrations is very much in the centre of attention. Input from other initiative, e.g. the EU Copyright work TDM (Text and Data Manual) can nicely link to the DS Curriculum. He established, 25 years ago, a one-time-charge license agreement with Verisign and Springer for life and at 95% discount. Most of the presenters defended a **neither top-down nor bottom-up but transversal approach** to cover all disciplines that deal with data, not only Computer Science. His team collaborate with CERN in the Open Aire project for publishing (contact is Salvatore Mele from the CERN library). Marek is an important player in the European Universities Association (EUA) and feeds the outcome of these discussions into their meetings.

- Thomasz Szapiro (ex-physicist, currently economist), spoke about the Conference of Rectors of Academic schools in Poland, a commission of accreditation and ranking, also covering budget flow from the ministry to the universities. The Polish Accreditation committee is to advise on the Data

Science Joint Programme on the question: *what is the end product*: a certificate of compentence? who designs/issues such certificates? who defines the quality measures?

- Dominik Batorski organises very well-attended DS meetups. 90% of the participants work in companies, hardly any academia. The DS experts show what they can do and employers come to seek talents. There are also other *meetups* on hadoop and R users, AI, ML, Big Data etc. Professional who need/want to complement their education have difficulties to do it in their spare time and own expenses and find the education programme relevant to their professional needs.

- Other polish (ICM) participants were Lidia Stepinska, studying the sociology of new professions and Marta Hoffmann-Sommer, ex-molecular biologist giving short seminars to PhD students to teach them how to put order to their data. They find the inter-disciplinary aspects especially enriching.

- Academic establishments come up with *Master's degrees in Data Science* every day now and the EDISON project would like to raise awareness of the multi-facetted aspects of the DS skills. Indeed, at the EPFL Research Day on June 8th, the IC Dean (Computing) James Larus announced a new Masters' programme in Data Science.

- Kevin Ashley from the Digital Curation Centre in Edinburgh represented the European Open Science Cloud pilot EOSCpilot. The pilot is responsible for the work package *Skills and Capabilities*. It is considered a *training* work package but the team tries to prove one acquires skills in multiple ways, e.g. from exchanges, such as an exchange with University of Amsterdam students, taken on June 1-2 to Edinburgh to work together with arts and design specialists on data visualisation.

- Sergiy Syrota, Ukranian Applied Maths professor from the National Technical University of Kiev, reported on their programme that could be decided without government approval, thanks to the establishment's prestige. Their naming problem is that *Business Analytics* are more popular than Applied Mathematics. They held a workshop to find a native language term for 'Data Science' and selected DATISTIKA. They are in touch with the European Consortium for Maths in Industry (ECMI), which establishes standards of maths in industry and try to get the DS jargon in the standards.

- Brenda Quismorio from Analytics' Association of the Philippines, represents ASEAN-5, a consortium of academic institutions from Singapure, Indonesia, Thailand, Philippines. Academic institutions can't catch-up with industry. The youth in those countries desires to follow high-level studies but unemployment is not solved by education. The reason is that a lot of companies outsource their processes to the Philippines but count on Artificial Intelligence and other automated processes only. Data Science skills will allow the workforce to offer an added value via the correct *interpretation* of the data. There are only 6 out of 2,000 universities in the Philippines that deliver DS undergraduate level degrees. Additional notes by Brenda: *The ASEAN - 5 collectively represents one of the most dynamic and promising growth regions with an estimated yearly GDP growth of close to 5%. They employ a variety of growth strategies: Singapore goes for innovation driven strategy; Malaysia and Thailand move towards technology and knowledge intensive industries; Indonesia and the Philippines maintain their competitive advantage in low skill industries. The achievement of their growth targets will highly depend on their abilities to equip their workforce with the needed skills. Interestingly, amidst the diversity, they share a number of skills challenges. Arguably, the most daunting challenge common to them is the inability of the educational institutions to meet industry requirements. The Philippines' major causes of these skills challenges are: weakness in STEM education, outdated curriculum, poor quality of instructors with insufficient industry experience, students lacking in industry exposure, inadequate instructional facilities, etc. Key to addressing these skills gap is the alignment among stakeholders i.e., employers, academia and government. Specific to addressing the shortage in analytics skills, a consortium i.e., the Analytics Association of the Philippines (AAP) was launched last May 24, 2017. Aside from linking and aligning the stakeholders, AAP aims to provide the analytics framework for the development of curricula, training modules and hiring talents,*

provide industry benchmarks, and thereby boost the analytics services provided by local or foreign companies. For its analytics framework, AAP is trying to integrate existing framework such as EDISON and APEC's Project DARE (Data Analytics Rising Employment). A number of AAP members took part in APEC's DARE Advisory group which convened in Singapore last May to draft the 10 DSA Competencies for the APEC region. In addition, there are currently very few universities offering a number of DSA programs in the Philippines. This is an indicator that there is much more to be done on the supply side. However, the good news it that the Philippines is putting the pieces together via AAP.

- Wout Los, University of Amsterdam and research infrastructures' specialist proposes the organisation of national workshops, so that the EDISON experience helps countries to create their own educational programmes in the DS field.

- Sue Geuens, President of http://www.dama.org/ is working on the *Data Management Body of Knowledge* - due at the end of June. The data governance contains all buzzwords of today and the policies and their inter-connection. The 2012 Harvard Business Review article: The Data Scientist: The Sexiest Job of the 21st Century is not anymore. People are being fired for having claimed competencies in Data Science, which they didn't actually have. In 2015, the 1st EU association for Data Science was organised. Other bodies, like the General Data Protection Regulation (GDPR, basically defending data ownership by the individual) are popping up and the professionals have to be aware.

EDISON meeting - Krakow 2016/09/27

Highlights relevant to CERN

Given CERN's experience with large data production, analysis and analytics, Maria Dimou suggests to:

- Find reviewers in the lab of the 4 main project documents
- Publicise the (imminent) survey of understanding of the data scientist's job
- Create a Vidyo room for holding meetings remotely with the main project players.
- Establish a collaboration with the Research Data Alliance (RDA) because of their global education mission.
- Check the portal and read about its functionality in the the CERN Edison twiki.

In more detail

The meeting agenda is attached to this twiki. Maria's free-format notes from the discussion:

The goals of the 27/9/2016 meeting were to collect participants thoughts on:

- the emerging components of the EDISON Data Science Framework (EDSF) - the four documents discussed
- how to best maximise the exploitation of the EDSF particularly in the context of European Research Data Infrastructure.

The full Expert Liaison Groups (ELG) members' list is http://edison-project.eu/edison/expert-liaison-groups-elg/elg-membership . CERN is part of the Employers group.

The documents discussed were:

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

These documents classify data specialists in Information and Communications Technologies (ICT) in 'families'. The job definition business reminds of the CERN benchmark jobs but the families are much greater in number as there are many sub-divisions.

- One commercial partner of the project participated, Jasper de Vries from company Kadenza in .nl. Consultant of many companies dealing with data, e.g. insurance companies.
- Donatella Castelli from the National research council in Italy.
- Community Portal being developed in Rome for the post-EDISON era, as the project is normally finishing next year.
- The Data Scientists (DS) are prepared via various channels, not only universities and company trainings but also self-education via MOOCs and other online courses.
- There is an issue at this moment about "WHAT is a DS". Hence, the project aims at building the job profile and the career path. Also to help universities to build a COMMON portfolio for accreditation and certification.
- Champion universities are Southampton, Perugia, Frankfurt, Luzern and Bradford. Next conference of champion universities is next February 2017 in Madrid. EDISON tries to increase the number of participating universities.
- Part of the project's challenge is to formalise the definition of people who inevitably move from an individual farm model in terms of data use to a supermarket model, i.e. pool data together and use across disciplines.
- Kathrin Beck - computational linguist (natural language recognition) in Max Planck Institute in Munich. Research Data Alliance (RDA) Global initiative representative. Total number of RDA members 4345 from 111 countries. Started 6 years ago. A lot of commercial companies are interested and government agencies wish to encourage participation because of the RDA dynamic explosion. EU-funded for Europe. USA and Australia also active. Organising training events, webinars, f2f training courses, all free of charge. People speak about what they do in their work and how this can be useful to others. They also have an Atlas of Knowledge (AoK). **They wish to share their educational material with others. Also, to UNIFY the definition of metadata, not so much for searching purposes but also for archiving. It will mean standardisation of nomenclature for volumes. SEEKING collaboration with other training providers.**
- Prof. Dimitar Trajanov - Saints Cyril and Methodius Skopje university. Data science is not a separate faculty. Machine learning and some web-based course material is the closest they have. They also have some students' projects, indirectly merging data from different disciplines, e.g. food - drug interactions. Their research results were fed into the Global Open Drug Data platform (GODD). They are now using the google BigQuery in Education for introductory classes in Data Science and Data Analytics. Data Visualisation and language processing for non-english speakers is a challenge. Very important to deal with "messy data", not clean datasets, as the commercial world is full of data to parse that are *not* uniform.

- Presenters from the EDISON project team were Steve Brewer (univ. of Southampton), Malgorzata Krakovian (EGI Foundation) and Yuri Demchenko (univ. of Amsterdam), the main editor of the 4 documents. They explained the distinction between Certification (recognition of an individual) & Accreditation (recognition of a school). The EU, companies' HR depts and other standardisation bodies are approached by EDISON to get acceptance of the Framework Definition. It is very important to know how to deal with "messy data", not clean datasets, as the commercial world is full of data to parse that are not uniform.

- Sustainability plan (Themis Athanasiadou- EGI Outreach): Chicken-&-Egg problem: To establishing

the EDISON brand is a prerequisite for MOOC providers to register their courses with EDISON. EDISON does provide Intellectual, human, virtual and physical resources. In order for the project to gain money to maintain up-to-date the Body of Knowledge and Data Scientist's profile definition, some services should be provided. These services are the 'registration' of external courses with the EDISON quality warranty. Other service is the portal with virtual labs on competence benchmark for data scientists to play before a job interview. This portal is developed in Rome by the italian project partners. When one takes the competency test on the portal, if found lacking some area of expertise, the portal offers the right online course and prompts the user to enroll. Parallel to EDISON activities: IBM Workbench and many more known in the Netherlands.

- Yuri Demchenko (EDISON papers' editor): After defining the Competence Framework we have to build an Online Educational Environment. It is worth the effort. The companies, like IBM, put 30% of their resources in education, they organise courses for which they charge 3K $ per person, so, the field is booming. There is a European Commission 'directive' for national initiatives in introducing digital education at all levels of schooling.

Various deadlines:

- European Commission Programme funding activities that bring research results into the market (FTI) Fast Track to Innovation 25/10/2016,
- Erasmus+ March 2017.
- EDISON2 29.03.2017. Also envisaging crowd funding.

A.O.B.

Yuri Demchenko said that Pascale Goy from CERN HR L&D gave a talk last March at EMBL in the UK about CERN planning to re-train the personnel. Slides attached to this twiki.

-- MariaDimou - 2016-09-29

This topic: ELearning > Edison
Topic revision: r13 - 2017-06-24 - MariaDimou