H2020 INFRASUPP-4 CSA Project

EDISON – Education for Data Intensive Science to Open New science frontiers

Yuri Demchenko
University of Amsterdam

Outline

- Consortium members
- EDISON Project Concept and Objectives
  - Challenges
- Project structure – WPs and Tasks
- EDISON Liaison Group and working with communities
  - EDISON workshops
- Next steps
- How to contact EDISON project team?
  - Project coordinator - Yuri Demchenko <y.demchenko At uva.nl>
  - EDISON project invites wide cooperation and contribution to Data Science profession definition
  - Join RDA Interest Group on Education and Training on Research Data Training (IG-ETRD) - https://www.rd-alliance.org/node/971
    - Join IG-ETRD mailing list <rda-edu-ig@rda-groups.org>

EDISON Consortium Members

Universities

1. **University of Amsterdam (UvA) - Coordinator**
2. University of Stavanger (UiS)
3. University of Southampton (UK)

Community related partners

4. European Grid Initiative (EGI)
5. FTK – Research Institute for Telecommunication and Cooperation (DE) and APARSEN (Alliance Permanent Access to Records of Science in European Network)

Industry partner

6. Industry: Engineering (Italy)

SME

7. InMark (Spain)

Project Objectives

Objective 1: Promote the creation of Data Scientists curricula by an increasing number of universities and professional training organisations.

Increasing the number of universities offering Data Science programs is essential in order to respond to the increased demand of the European research, industry and public sectors.

Objective 2: Provide conditions and environment for re-skilling and certifying Data Scientists expertise to graduates, practitioners and researchers throughout their careers.

To provide conditions for a sustainable increase in the number of Data Scientists in Europe, EDISON will also support the training by professional organisations of the self-made expert, in the light of reaching the target number of graduates that gradually will satisfy the level of research and industry demand in coming 5-8 years.

Objective 3: Develop a sustainable business model and a roadmap to achieve a high degree of competitiveness for European education and training on Data Science technologies, provide a basis for the formal recognition of the Data Scientist as a new profession, including supporting the future establishment of formal certification for practicing “self-made” Data Scientists.

Objectives 1: Data Science model curriculum

Objective 1: Promote the creation of Data Scientists curricula by an increasing number of universities and professional training organisations.

Increasing the number of universities offering Data Science programs is essential in order to respond to the increased demand of the European research, industry and public sectors.

- A general model/framework of the required competences and skills starting from needs of existing and emerging e-Infrastructures, different scientific domains and industry sectors (Competence Framework for Data Scientist - CF-DS) including Glossary and Taxonomy of competences and skills.
- The Body of Knowledge (BoK) for the Data Scientist Profession (DS-BoK) that will be used to map required competencies/skills and existing academic, research and technology disciplines
- A Model Curriculum for Data Science (MC-DS) that will provide a reference implementation of the proposed CF-DS and DS-BoK.
- A promotional and networking framework (called EDISON-Net) that establishes communication with top-level European institutions in European countries

Objective 2: Certification and re-skilling

Objective 2: Provide conditions and environment for re-skilling and certifying Data Scientists expertise to graduates, practitioners and researchers throughout their careers.

Provide conditions for a sustainable increase in the number of Data Scientists in Europe, in the light of reaching the target number of graduates that gradually will satisfy the level of research and industry demand in coming 5-8 years.

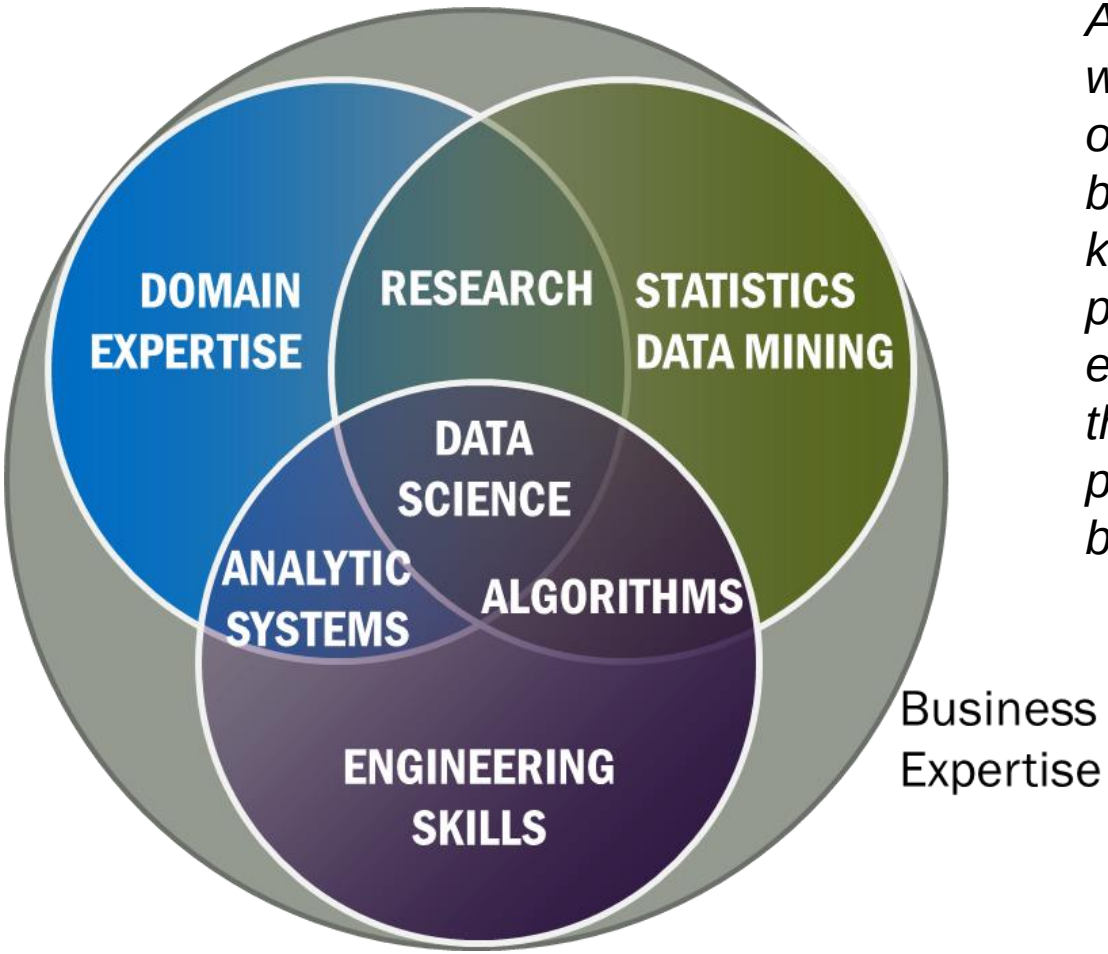
- Seamless access to Education and Training e-Infrastructures, to support specialist training on Data Intensive Science and technologies that will include access to data sets for educational purposes, data management and analytical tools and cloud based virtual labs and classrooms.
- The creation of an EDISON Marketplace of existing and developing education and training courses, education materials and other resources, by leveraging on the EGI Knowledge Commons and the Training Market initiative.
- A model and a framework for Data Science education programme accreditation and professional certification, also targeting practicing “self-made” Data Scientists that want to obtain a formal certification via an academic or professional training institution.
- A network of experts, champion universities and professional training organisations and other stakeholders that will ensure consistent definition of the CF-DS, the DS-BoK and the MC-DS by sharing their experiences.

Objective 3: Sustainability model

Objective 3: Develop a sustainable business model and a roadmap to achieve a high degree of competitiveness for European education and training on Data Science technologies, provide a basis for the formal recognition of the Data Scientist as a new profession, including supporting the future establishment of formal certification for practicing "self-made" Data Scientists.

- An education model suitable for all major learning paths (piloting cases) and stakeholders for education and training in Data Science from the consumer, operator and professional community sides that will include multiple learning models (e.g. residential, lifelong learning, e-Learning) that will support Data Science education evolution and individual career path management as technologies evolve.
- A business model that incorporates the proposed education model into a sustainable cycle of the research data value chain including the major stakeholders and actors from research, industry and government; produce a roadmap for a sustainable education and training services as a part of overall data driven technologies development in Europe.
- Establish EDISON Liaison Group(s) (ELG) of independent experts that represent the major stakeholders in Data Science that will work as a consulting body for the project and will create a basis for future independent expert group for universities and for EC.

Data Scientist skill definition

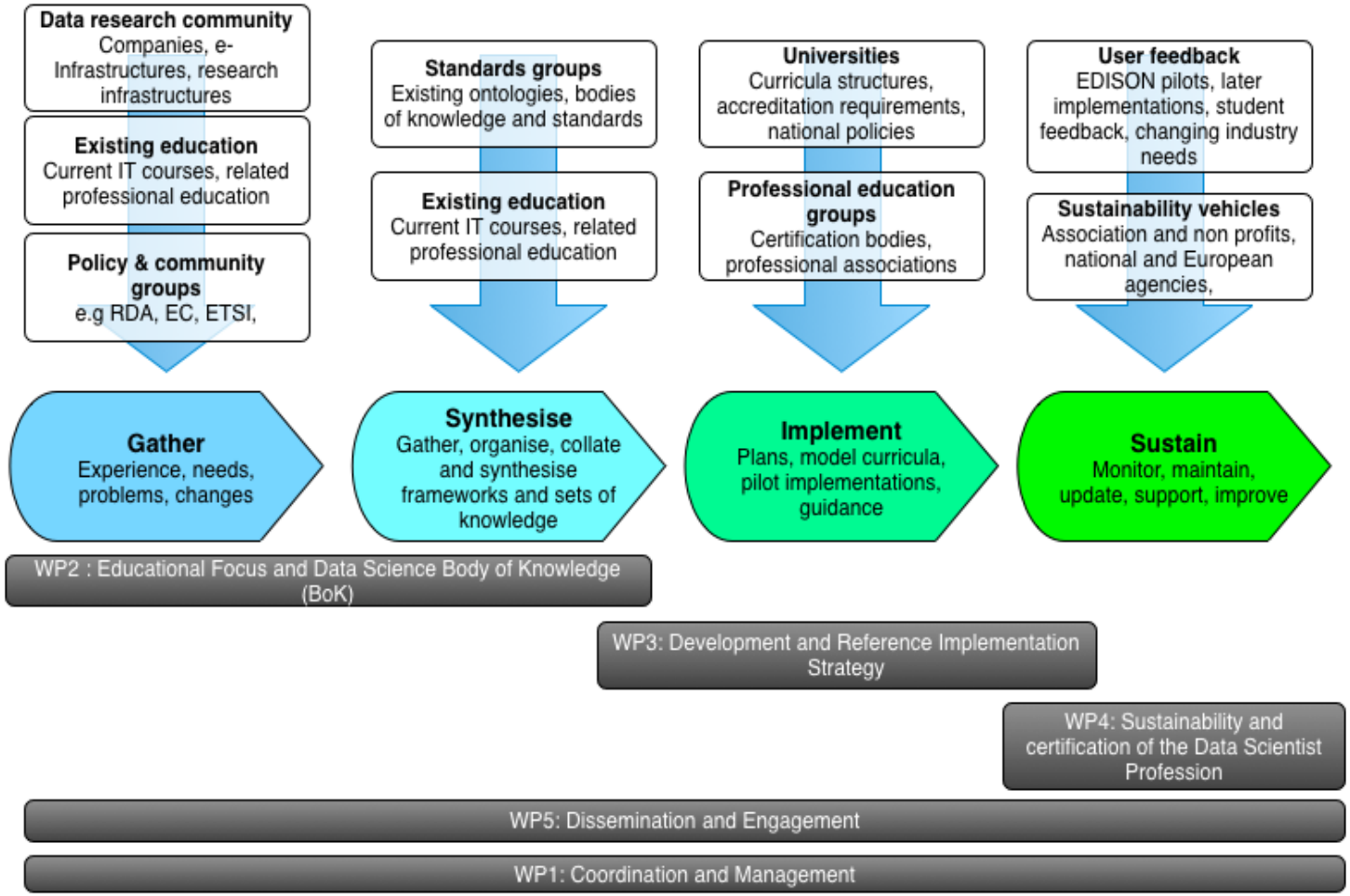


*A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the big data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

Gaps: Education and Training for Data Science and e-Infrastructure

- Universities both in Europe and in USA (worldwide) don't offer sufficient possibilities for educating new type of specialists.
  - There are few Data Science programs offered in Europe and US.
  - Some of currently offered programs are done based on slight modification or simple re-branding former curricula on Business Analytics, Data Analytics or Machine Learning.
  - Existing academic disciplines and education programs can create a strong basis for Data Science education and profiling
- There is no common vision and model how to combine theoretical and practical components in university curricula and create a necessary education and training infrastructure for practical skills acquisition
- There is no (limited) coordination between industry and academia to achieve sustainability of the new specialists education and training that could avoid current competition between industry and research for new skills and competencies
- The new profession of the Data Scientist is not well defined and not recognised as a distinct profession what leads to not consistent education and training programs and not correct expectation from employers and job agencies

Basic methodology of EDISON, inputs and work packages



WPs Structure

WP1 Coordination and Management

WP2 Educational Focus and DDS-BoK

- T2.1 Education and training needs and competencies
- T2.2 Inventory Education and Training resources
- T2.3 Data Scientist: DS-BoK and CF-DS profiles

WP3 Development and Reference Implementation

- T3.1 Development of the EDISON MC-DS
- T3.2 Piloting Use Cases A and B: Data Science Education and Training
- T3.3 Use Case C: Course and Programme accreditation
- T3.4 EDISON Online Educational Environment

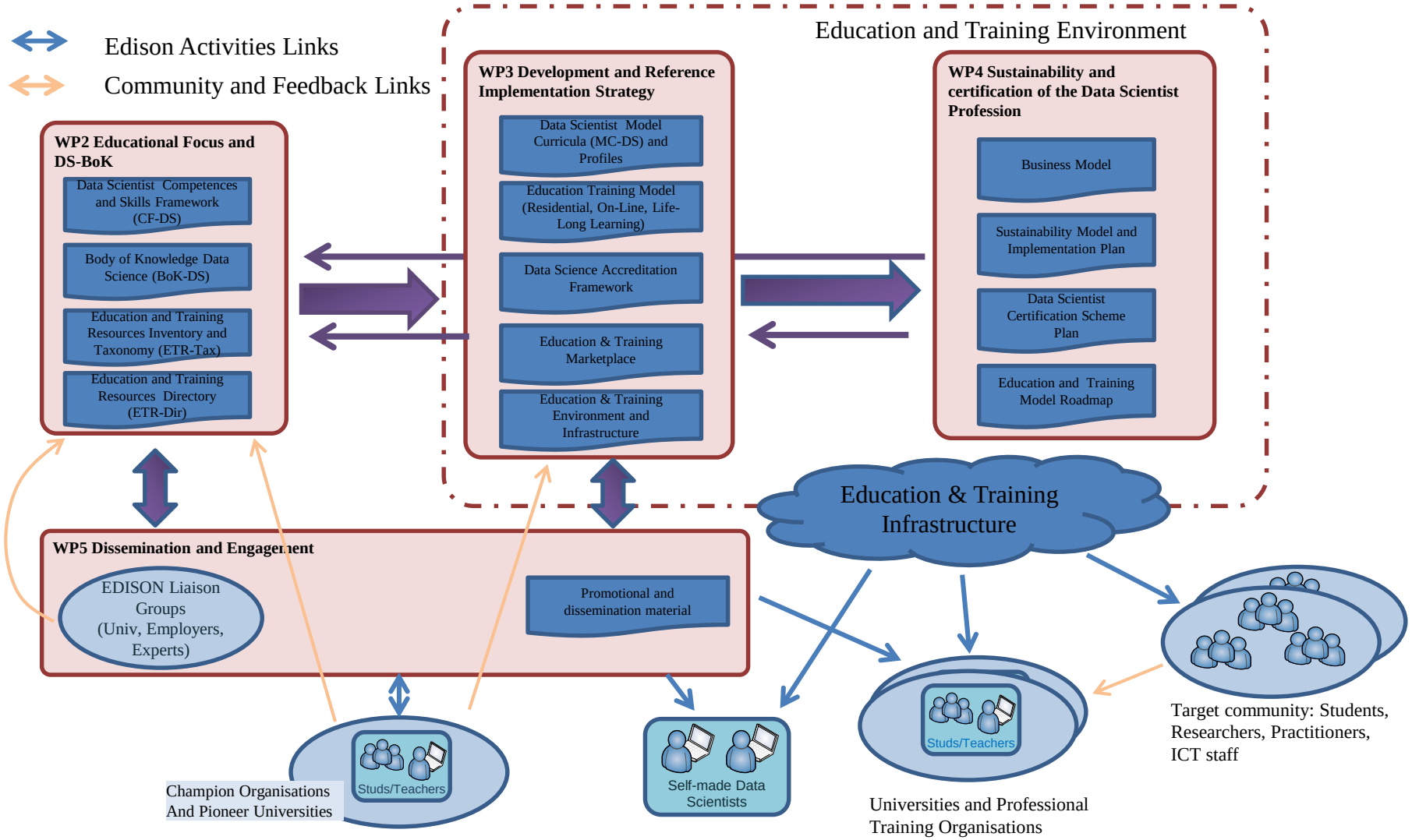
WP4 Sustainability and certification of the DSP

- T4.1 Strategic Market analysis for sustainability
- T4.2 Business models definition
- T4.3 Sustainability: Plan and implementation
- T4.4 Definition of a certification scheme

WP5 Dissemination and Engagement

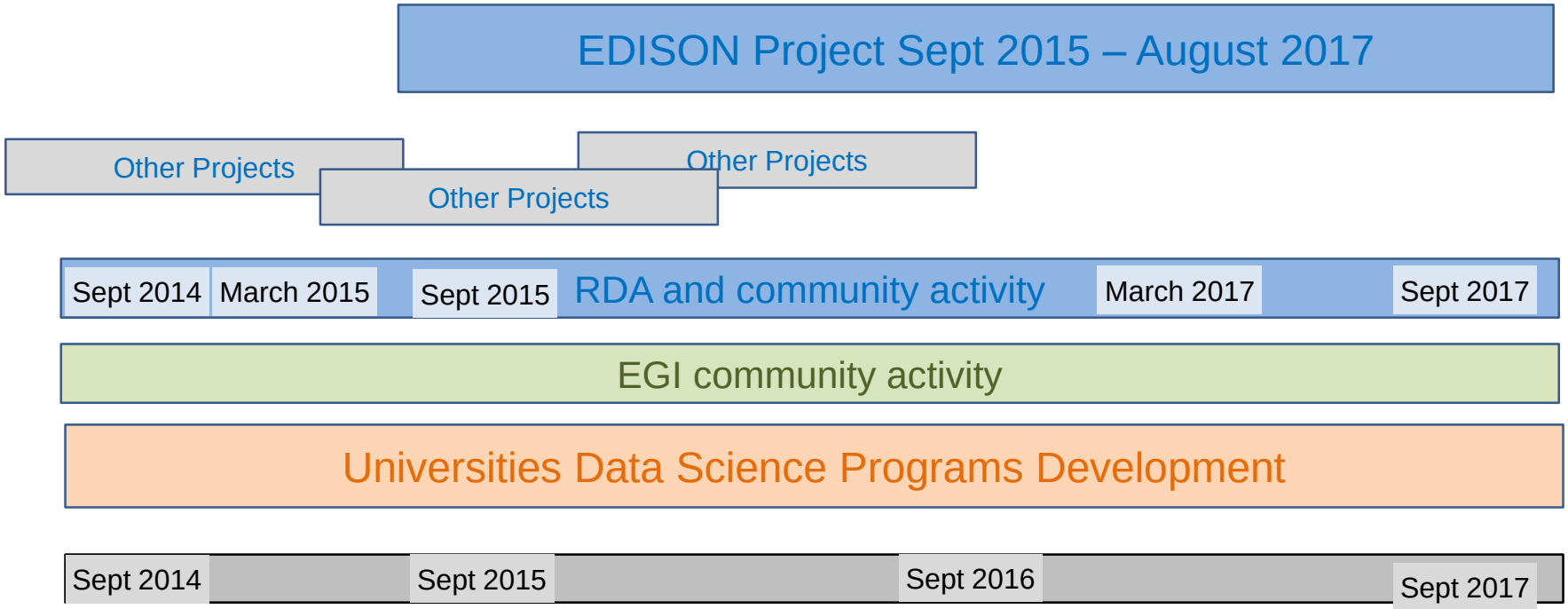
- T5.1 Management of the DEP
- T5.2 Production and distribution of dissemination material

WPs interaction and main tasks



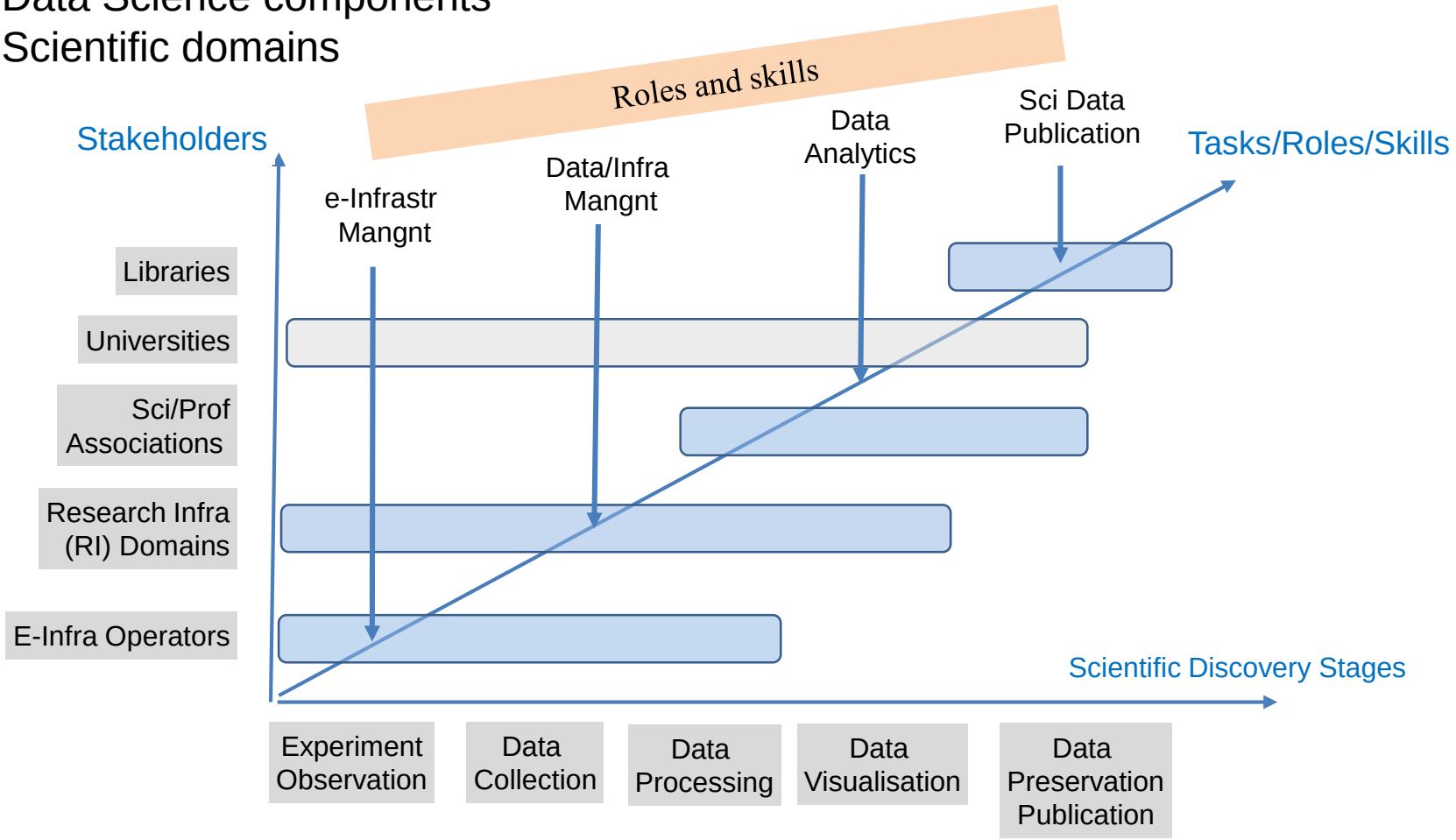
Ongoing Education Programs Development Timeline and EDISON Project Lifespan

- The proposal involves partners that run individual educational and training programs development
  - Enthusiastic universities and community driven activities
  - EGI training program/marketplace development
  - RDA Interest Group on Edu & Training, IG on Big Data Infrastructure, WG on Big Data Analytics and others
- **Funding is requested for coordination of these activities and building sustainable Data Science education environment and infrastructure**

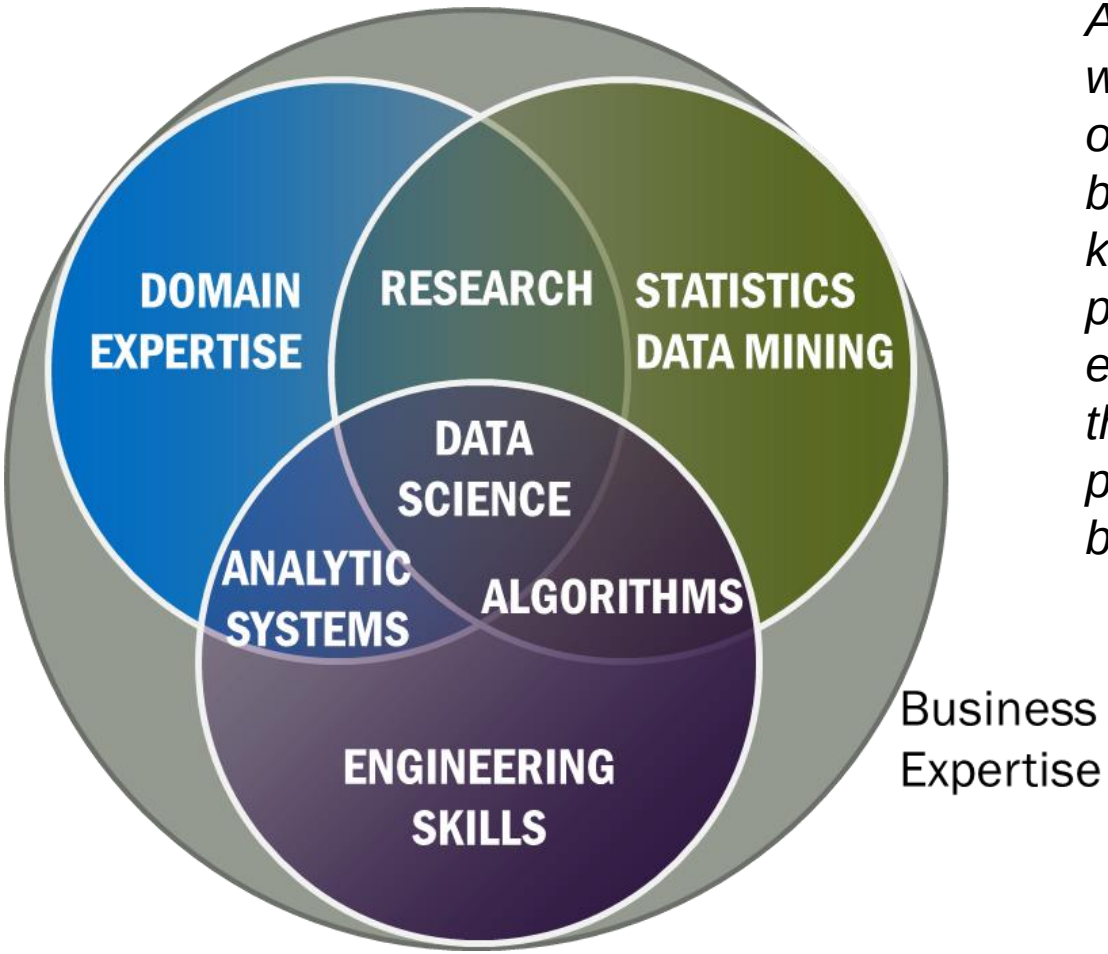


EDISON Dimensions: Stakeholders, Scientific Discovery Stages, Data Management Functions

- Partners with experience
- Data Science components
- Scientific domains



Data Scientist skill 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.*

Data Science Curriculum Definition

Education Infrastructure/Environment

- Skills and competencies
- Common Body of Knowledge (CBK)
- Stakeholders
- Training/education infrastructure
- Datasets
- Approaches: Bottom up vs Top down

Specialisations

- Data Analytics and Tools
- Scientific Experiment Automation and Scientific Data Infrastructure
- Profiling for scientific domains
- Data Archiving, preservations, Librarian

Example CBK Master Program

1. Big Data Definition and Big Data Architecture Framework, Data driven and data centric applications model, Stakeholders and Roles
2. Big Data use cases and application domains taxonomy and requirements, Big Data in industry and science
3. Data structures, SQL and NoSQL databases
4. Data Analytics Methods and Tools, Knowledge Presentation
5. Big Data Management and curation, Big Data Lifecycle, Data Preservation and Sharing, Enterprise Data Warehouses, Agile Data Driven Enterprise
6. Cloud based Big Data infrastructure and computing platforms, Data Analytics application and new Data Scientist skills required
7. Computing models: High Performance Computing (HPC), Massively Parallel Computing (MPP), Grid, Cluster Computing
8. Big Data Security and Privacy, Certification and Compliance

Education&Training Programs Development Approaches

- Reuse experience from the existing related program
  - Cloud Computing Technologies and Tools
  - Theoretical Informatics, Data Analytics and Artificial Intelligence
  - Collect examples/projects and do taxonomy
  - Science/subject domain and “long-tale” science
  - Libraries and data archives
- General (inter)active education principles
  - Common Body of Knowledge
  - Bloom’s taxonomy
  - Pedagogy vs Andragogy
  - MOOC: possibilities and limitations
  - Discussion forum for online education, and Research and Reading assignments for on-campus education

How to contact EDISON project team?

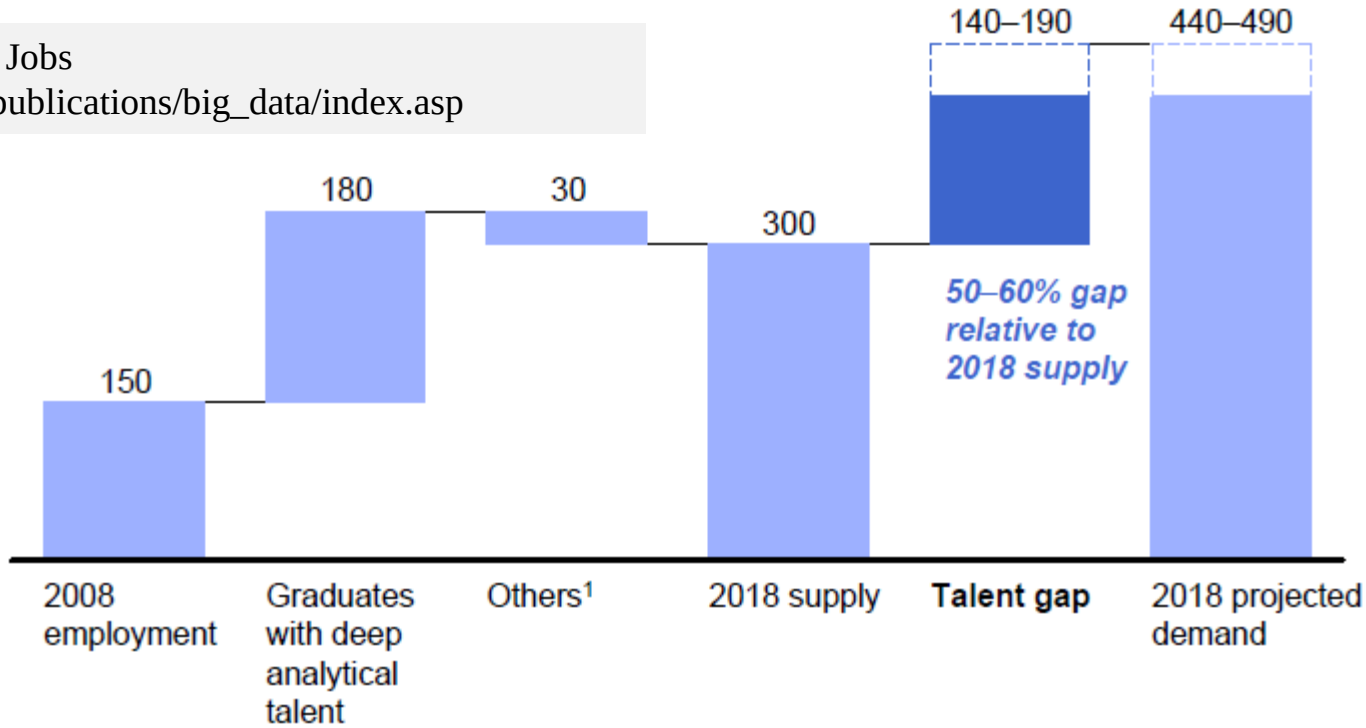
- EDISON project invites wide cooperation and contribution to Data Science profession definition
- Project coordinator – University of Amsterdam
  - Yuri Demchenko <y.demchenko At uva.nl>
  - Join RDA Interest Group on Education and Training on Research Data Training (IG-ETRD) - https://www.rd-alliance.org/node/971
    - Join IG-ETRD mailing list <rda-edu-ig@rda-groups.org>

Additional Information

- McKinsey Institute on Big Data Jobs
- Strata Survey Skills and Data Scientist Self-ID
- Big Data technology domain definition

Data Scientist: New Profession and Opportunities

McKinsey Institute on Big Data Jobs
http://www.mckinsey.com/mgi/publications/big_data/index.asp



- There will be a shortage of talent necessary for organizations to take advantage of Big Data.
  - By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as
  - 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions

SOURCE:US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey analysis

Strata Survey Skills and Data Scientist Self-ID

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Analysing the Analysers. O'Reilly Strata Survey – Harris, Murphy&Vaisman, 2013

- Based on how data scientists think about themselves and their work
- Identified four Data Scientist clusters

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

Skills and Self-ID Top Factors

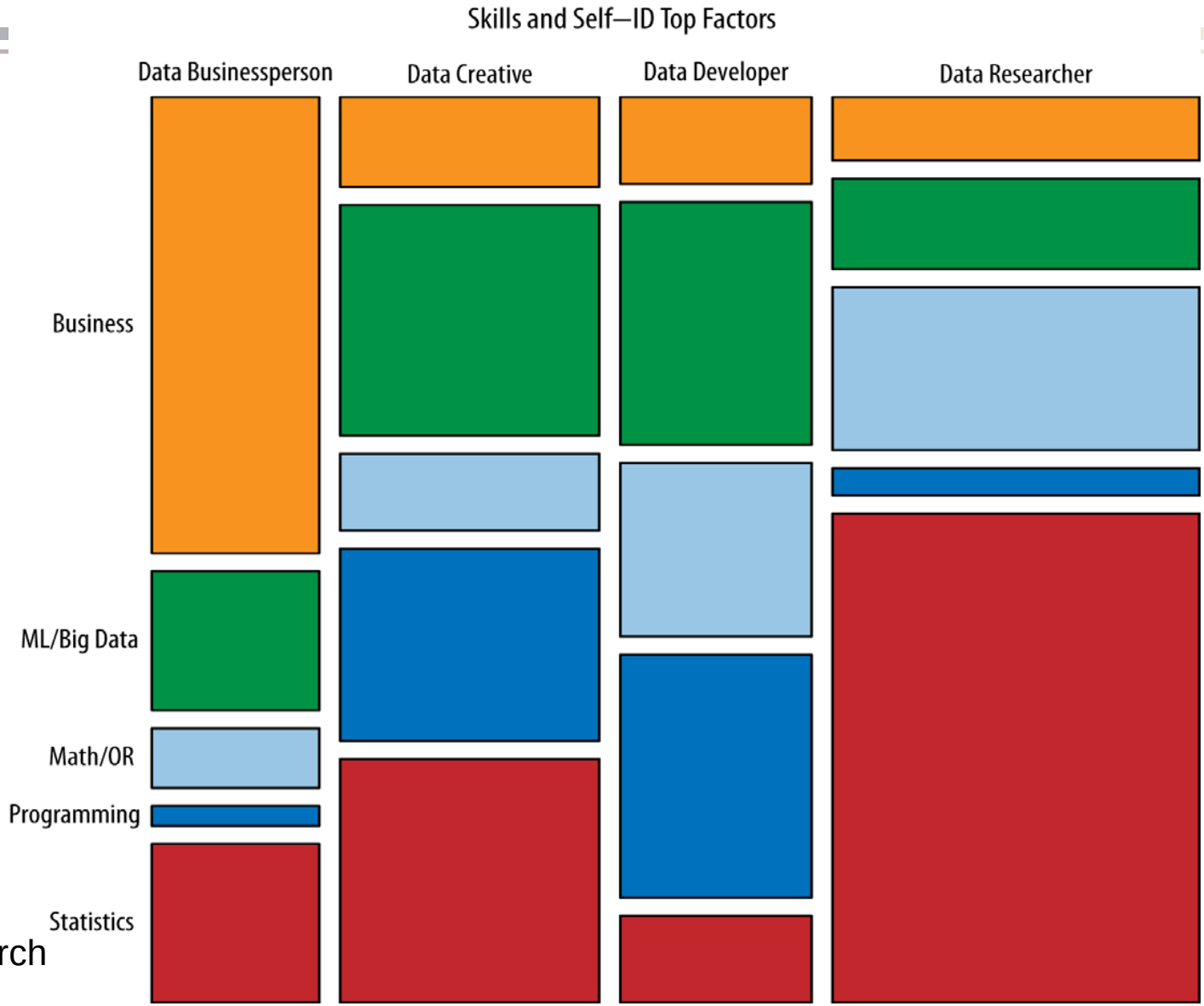
Business

ML/BigData

Math/OR

Programming

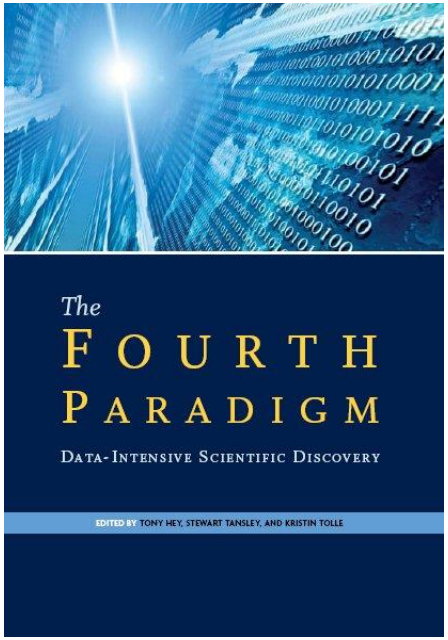
Statistics



ML – Machine Learning

OR – Operations Research

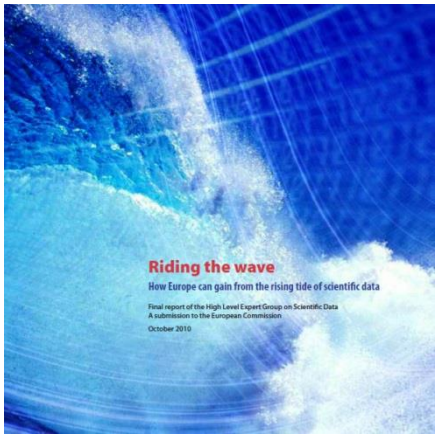
Visionaries and Drivers: Seminal works and High level reports



The Fourth Paradigm: Data-Intensive Scientific Discovery.

By Jim Gray, Microsoft, 2009. Edited by Tony Hey, 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

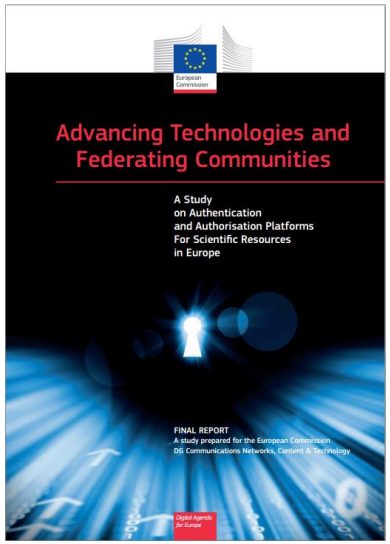
RDA RESEARCH DATA ALLIANCE

Research Data Sharing without barriers

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

NIST Big Data Working Group (NBD-WG)

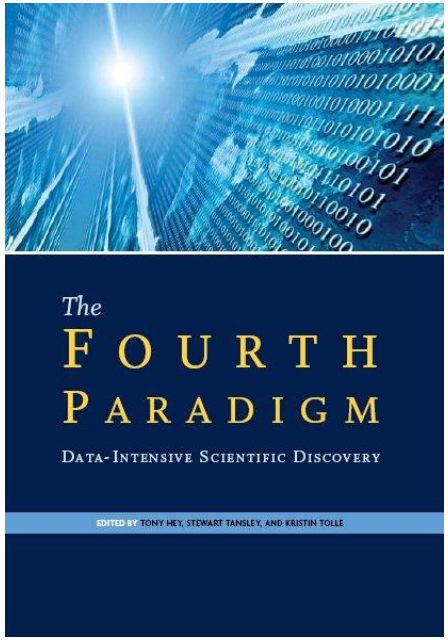
https://www.rd-alliance.org/



AAA Study: Study on AAA Platforms For Scientific data/information Resources in Europe,

TERENA, UvA, LIBER, UinvDeb.

The Fourth Paradigm of Scientific Research



1. Theory and logical reasoning
   - Based on observation and *imagination* of ancient philosophers
2. Observation or Experiment
   - E.g. Newton observed apples falling from the tree to design his theory of mechanics, verified with other observations, validated with experiments
   - 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 Intensive Science)
   - e-Science as computing and Information Technologies empowered science
   - More data beat hypnotized theory
     - But not to use “microscope blindly”