PDENG PROGRAM DATA SCIENCE

The Professional Doctorate in Engineering program Data Science aims to educate students with an academic master degree to become professionals, who are of direct value for industry and business, or become the innovative entrepreneurs in the data science ecosystem. The Mariënburg Graduate School Data Science embedded in the Jheronimus Academy of Data Science in which the universities of Tilburg and Eindhoven cooperate offer true opportunities and create excellent conditions to accomplish this objective. Within the Mariënburg infrastructure, the PDEng program may become part of a data science community, where postgraduate education, innovation and entrepreneurial activities (start-ups), lifelong learning, and scientific research are integrated. Students are employed by the Eindhoven University of Technology as Technological Designer in Training.

Data Science

Data Science can be broadly defined as the study and design of computational principles, and automated methods and systems, to analyze massive and complex data to extract useful information. As such Data Science lies at the crossroads of computer science, applied mathematics, and statistics. Large data sets are now generated by almost every activity in science, society, and commerce — ranging from molecular biology to social media, from sustainable energy to health care. Data Science asks: *How can we efficiently find patterns and the dynamics of these patterns in these vast streams of data in the context of the data environment?* Many research areas have tackled parts of this problem: machine learning, and data and process mining focus on finding patterns and making predictions from data; databases are needed for efficiently accessing and integrating data and ensuring its quality; algorithms and architectural models are required to build systems that scale to big data streams; natural language processing, computer vision, and speech processing are each needed for analysis of different types of unstructured data. Knowledge of law and ethics are required to understand legal and ethical implications; knowledge of management and business to enhance market decisions and explore economic consequences of choices. Recently, these distinct disciplines have begun to converge into a single field called Data Science. Data Science becomes a new frontier for design.

> *"A data driven organization acquires, processes, and leverages data in a timely fashion to create efficiencies, iterate on and develop new products, and navigate the competitive landscape."*
>
> *Patel, president LinkedIn*

Bridge to a career in Data Science at professional doctorate level

Top companies, markets and businesses in many fields are hiring big data infrastructure professionals at doctorate level to help them store, process, and make accessible the terabytes of data that they collect every day. While the amount of data produced and stored is growing exponentially, there is a severe shortage of talent to design, develop, maintain, and optimize the infrastructure and data pipelines necessary to efficiently store, analyze, and extract the valuable insights from data. The focus of the PDEng Program Data Science (DS) is on the design practice and methodology within the Data Science field. The program offers an interdisciplinary learning environment devoted to professional training in competencies and skills that are requested from a professional data scientist, thus integrating the fundamentals from analytics, computing, data and process mining, and business intelligence, from a design perspective. The intent of the program is to broaden and specialize the professional profile of its students in a personalized way. Candidates for the program are graduate students with a master degree in fields such as engineering, econometrics, mathematics, computer science, geophysics, bio-informatics, and computational chemistry, aiming at a professional career in a big data environment that truly leverages their academic experience and talents in the fast-growing, in-demand field of Data Science.

Data Scientist & Data Engineer

The amount of data produced across the globe increases exponentially and will continue to do so in the foreseeable future. In business institutes, internet markets, and industries, servers are overflowing with usage logs, message streams, transaction records, sensor data, business operation records, and mobile device data. An efficient analysis of these huge collections of data — big data — will create significant value for any economy by enhancing productivity, increasing efficiency, and delivering more value to consumers. Studies estimate that trillions of euros of value in efficiency improvements and economic growth can be unlocked by extracting actionable knowledge from the deluge of data now being collected in almost every sector of the economy. Nowhere has the benefit of analyzing data been felt more strongly than at top technology companies. Throughout the world, many Business Analytics companies are founded to support the acquisition, exploration and analysis of data, to create value from data to do so, company's first need to be able to reliably store, process and query its huge inflows. As a result, the data infrastructure needs to be distributed, scalable, and reliable, which is not a trivial engineering task given the terabytes of data involved. Currently, in organizations, besides the business analyst, one recognizes data engineers and data scientists. The ***data engineer*** creates and maintains the robust big data pipelines.

> *"We are on the cusp of a tremendous wave of innovation, productivity, and growth, as well as new modes of competition and value capture—all driven by big data as consumers, companies, and economic sectors exploit its potential," write the authors of Big Data: The Next Frontier for Innovation, Competition, and Productivity, a comprehensive research study published by the McKinsey Global Institute.*

The ***data scientists*** develops tools to analyze the data. Data scientists and data engineers form the basis of the data teams that have quickly become a central part of most technology companies' technical teams.

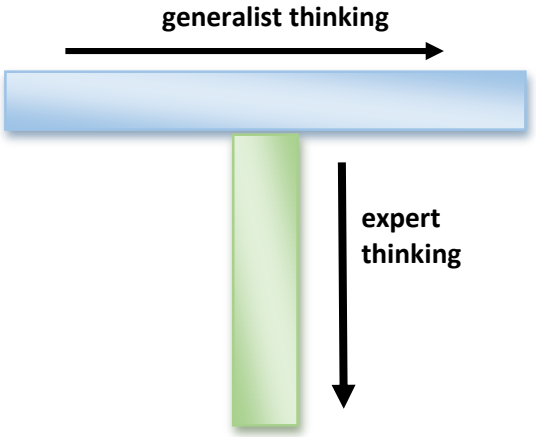
Roughly speaking, the data engineer combines skills from computer science and skills from statistics skills, with the emphasis on computer science skills; the business analyst combines domain expertise with skills from statistics with the emphasis on domain expertise; the data scientist combines skills from computer science, skills from statistics, and domain expertise. The data scientist oversees the full business-to-data-to-business value chain.

Data Science professionals

At present, data scientists and data engineers come from the traditional areas of computer science, mathematics, and engineering. They leveraged their underlying skills to enter this fast changing, dynamic area. This includes researchers or professionals coming from fields such as computer engineering, electrical engineering, embedded systems, network engineering and systems engineering. Individuals, who have studied mathematics and computer science fundamentals and have worked with large code bases or managed computational infrastructure in fields such as aerospace engineering, nuclear engineering, chemical engineering, mechanical engineering, physics and mathematics, have all successfully made the transition.
Engineers and PhDs coming from backgrounds like the ones listed above often take a long and winding road to get into the field, learning the tools used in big data informally over long periods of time and through chance encounters with the profession. While serendipity may be a good way for people to discover a field in its infancy, as the field matures and as demand grows, there needs to be a more direct and efficient route into the profession.
At this point, the PDEng program Data Science wants to step in and become the program par excellence to develop the next generation of leading Data Science professionals. The main challenge for Data Science is an educational one. There is an overwhelming number of tools, skills, and competences needed to survive in the Big Data ecosystem. Thus the ability to sense and act upon the knowledge dynamics effectively is a must-have for the Data Science professional that goes far beyond what can be expected from students at master level.

T – profile

Data Science professionals have interdisciplinary skills in analytics, computing, statistics, and business intelligence; have a creative, design oriented, problem solving attitude; have strong nontechnical skills related to project management, working in teams, communication, and presentation. A data science educational/training program draws from the strengths of a variety of disciplines, such as computer science, mathematical sciences, design science, and business management. Indeed, the PDEng program will deliver T-shaped graduates, who combine breadth (e.g., computer science, mathematics and statistics, business, and ethics and law) with deep knowledge and skills in resolving real-world data science problems. Only then students understand motivations and arguments of all people involved and recognize, what it takes to collaborate and deliver value. As generalist, they oversee the complete data value chain, as experts, they have detailed knowledge and skills in one of the design related subdomains of Data Science.



Candidate's profile

Candidates of the DS program show high motivation to learn combining generalist thinking and expert thinking, formulating opportunities and solving complex problems in a data driven environment. They possess an academic master degree showing a strong background in a combination of statistics, computer science, and mathematics. The candidates are entrepreneurial minded. Their profile can be typified as

- ✓ **Technical expertise**: the starting point is expertise in some scientific discipline at Master level, typically related to statistics, mathematics, and computer science
- ✓ **Curiosity**: desire to discover, distill, and model a problem down to a clear set of concepts and hypotheses that can be tested
- ✓ **Story telling**: desire to develop the skill to use data to tell a story and to be able to communicate it effectively
- ✓ **Cleverness**: desire to develop the ability to look at a problem in different, creative ways

Mission of the program

The two-year interdisciplinary PDEng program Data Science combines statistics, computer science, mathematics, and design theory with the business acumen to explore data sets, gather actionable insights, visualize results, and communicate meaningful findings taking into consideration underlying legal and ethical contemplations. Graduates make sense of data and have the ability to articulate their discoveries and recommendations to those not schooled in the world of data in the frame of industrial and business design and decision processes.
With strong industrial partners, students can be provided with the resources and opportunities to be engaged in purpose-driven training projects, formal course work, and extensive one-week workshops/assessments. Depending on their background students are offered a personalized educational/training program. The focus can be diverse ranging from general data analytics and computing, mathematical analytics, business analytics, to specialized concentrations in financial analytics, healthcare analytics, and analytics for sustainability.
Students in the program develop the generalist skills to have an overview of the full data chain and specialist skills in a Data Science area or data domain. They learn the design fundamentals of statistics and related statistical software; be grounded in data mining and process mining, and predictive modeling and simulation, in a design context; develop skills in the design of algorithms and their implementation, and use of programming; identify scientific problems and processes; and be capable of facilitating effective communication with both scientific and non-scientific collaborators. The program trains students to work in teams, communicate with professionals in the data domain field, carry out project management, and learn the social skills to become of added value to business and industry.

Besides its focus on the specialist skills, the program's aim is to educate students such that they are capable to:

1) independently acquire knowledge on an academic level;
2) to integrate and apply this knowledge for external stakeholders, who have domain expertise, but lack data science expertise;
3) to manage the design process with respect to project- and time-planning, taking the constraints of stakeholders into account;
4) to clearly communicate results, conclusions, and recommendations;
5) to coach others as participant in the process.

Some of these aspects are also addressed in MSc-programs; we require students of the PDEng program DS to perform in *all of these*, such that after graduation they can directly act as professional data scientist or data engineer for complex projects. Both the contents and the educational choices aim at enabling students to obtain a broad experience, and almost all aspects will be evaluated explicitly in each education component (see Appendix 1).

"The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that's going to be a hugely important skill in the next decades."

Hal Varian, Google Chief Economist

Skill development

A successful Data Science professional has technical and nontechnical skills at a high level and has integrated these skills optimally. Thus the program focuses on the ability to work with individuals from different disciplines within one team, the ability to give and receive feedback, to be organized, to be motivating, to listen and summarize, and to set up meetings. They all relate to nontechnical skills that are indispensable in the execution of any Data Science related activity.

The technical skill development in the program relates to

- Assessment of a data domain in light of data opportunity and data exploration, data ethics and law
- Design of algorithms and methods to collect, organize, and analyze data with use of state-of-the-art data analytic technologies in order to
  - Find and integrate rich data sources, and discover possible dynamics
  - Cleanse and process data so that it is consistent
  - Work with large volumes of data despite hardware, software, and bandwidth constraints
- Design of data mining and/or process mining application strategies
- Predictive model design
- Qualitative data analysis
- Application of information extraction techniques and design of data management
- Design of methods to present and visualize data
- Business analytics in relationship to the respective data domain

Data Science • Data acquisition and data management • Data exploration • Data and process mining • Predictive modeling • Data visualization	
Tool building • Data processing • Data cleaning • High performance computing • Algorithm Design • Optimization • Software Engineering	**Data domains** • Health • Energy • Smart mobility • Agro & Food • Financial Markets • Life Science • e-Marketing • Pharma • Accountancy, Fiscal and Legal

Professional doctorate

A candidate who is awarded the Professional Doctorate, Data Science PDEng is expected to have adequately developed the technical and nontechnical skills as mentioned in the previous paragraph. They have integrated these skills in the following training and learning outcomes:

- Systematically acquired an understanding of a substantial body of knowledge and experience with a feel for entrepreneurial issues and challenges that are at the forefront of the professional Data Science practices
- The ability to conceptualize, design, and implement a project for the generation of new knowledge, application or understanding at the forefront of Data Science disciplines; to adjust the project design in the light of unforeseen problems
- A detailed understanding of applicable Data Science techniques for design and analysis at academic level
- Ability to make informed judgments on complex issues in specialist fields, such as agro & food, health, financial markets and life science, often in the absence of complete data, and be able to communicate their ideas and conclusions clearly and effectively to specialist and non-specialist audiences
- The qualities and transferable skills necessary for employment requiring the exercise of personal responsibility and largely autonomous initiative in complex and/or unpredictable situations, in professional and equivalent environments.

Program training and educational elements

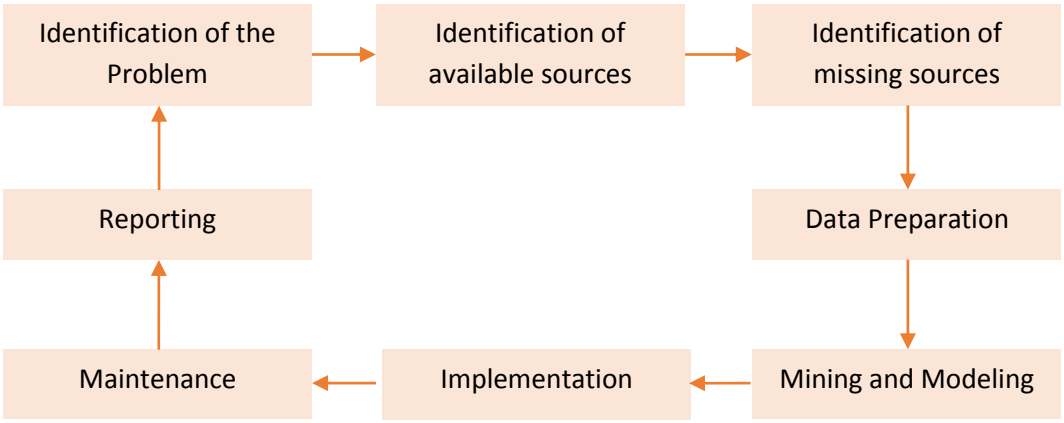
The two-year PDEng program is basically structured as follows. In the first year, students work in teams on real-world data science themes/data challenges in the scope of the data science modular build curriculum. In their second year, they carry out a 12 month full time data science related project in industry.

DS educational training program philosophy

In its education, the emphasis of the designer program is on the transfer of skills in relationship to knowledge rather than on transfer of knowledge and skills, separately. We refer to the skill to synthesize and solve, resulting in a design, that arises from an analytic approach towards a usually vaguely posed commercial, business or industrial problem, often within a technological context. To make a design and to do research are related activities, but the researcher is supposed to possess qualities different from the ones of the designer, and, also, to be led by a different motivation and personality. A design activity involves a design process with well-defined stages and competences that are of use in each of these stages. Likewise a research activity is determined by a research process with its corresponding stages. It is hard to define the clear distinction between design and research. Globally speaking research concerns the ‘why’ and design concerns the ‘how’.

Knowledge transfer in the program is based on the T-profile concept, with focus on broadening the students' knowledge of the data science landscape and specializing them on various aspects of data science on a tailor made basis. In this context, the essence is: (1) to be able to apply knowledge at hand to solve a problem, (2) to know what knowledge is of use in solving a problem, (3) to detect and acquire missing knowledge fast. These three learning aspects reach much further than what may be expected from a master student. They ask for maturity and responsibility of student's own learning process. They ask for lecturers with the right attitude and motivation. No doubt, the educational content of the program is closely related to the knowledge that the universities can offer as knowledge suppliers; as well, this content is determined by the clients, who want to apply innovating knowledge in their organizations.
By means of generalist courses the program takes care of the broadening of the knowledge of its students. In these courses the relationship and coherence between distinguished elements of the data cycle is provided. Apart from lecturers from the university, also lecturers from the ecosystem with a clear perspective on the Data Science field are involved. The modules concern the full business to data to business cycle with the emphasis on design aspects and integration of skills, in particular the technical and nontechnical ones. Although 'cycle' suggests a certain sequencing in the process, in a practical situation one jumps up and down the cycle. Thus, each distinguished part depends on all the others.

- *Identification of an opportunity or problem that can be dealt with by exploiting data*
  Assessment of expected business value (business includes institute, industry, business, government, foundation)
- *Identification of available data sources*
  Data quality, data ethics and integrity, data privacy, data storage and process, data exploitation in the business perspective
- *Identification of necessary additional data sources*
  Data storage, design of experiments, data accessibility, data costs, licensing
- *Identification of regulatory context, risk and compliance*
  Privacy & property rights, risk management, security measures
- *Data preparation*
  Data cleansing, data transformation, data exploration
- *Data analysis*
  Data mining, machine learning, statistics, process mining, time series analysis, multivariate analysis, network analysis, hypothesis testing, modeling from data, classification, data visualization
- *Implementation and development*
  FSSRG: Fast, Simple, Scalable, Robust, Generic
  Verification of the implementation, fine tuning, simulation set-up, key-performance-indicator extraction, assessment of predicted business value
- *Maintenance*
  Validation of the implementation, updating, and outsourcing, assessment of created business value
- *Communication of results*
  Verification/validation with domain experts, reporting (results, conclusions and recommendations), software documentation.



Before students can start a module, they are expected to have a basic level of knowledge and skills. If they have not, then they should do a homologation. Lecturers are expected to coach the learning process of the students rather than be the sole responsible for the content of knowledge transferred to the students. By this, the program encourages and trains its students in the capability to become self-directed learning professionals *) who take the responsibility to acquire knowledge provided by any medium, e.g., experts in the field and scientific literature. This marks a clear distinction with the learning processes offered to students at master level. Each of the modules address two fundamental questions: (1) what data science related (sub) problems can be solved with the knowledge and techniques clustered in the module? (2) What are the aspects corresponding to the design processes involved? To create real life conditions, theme projects and assignments from industry are the drivers in the process of knowledge transfer.

**) In its broadest meaning, 'self-directed learning describes a process by which individuals take the initiative, with our without the assistance of others, in diagnosing their learning needs, formulating learning goals, identifying human and material resources for learning, choosing and implementing appropriate learning strategies, and evaluating learning outcomes.* (Knowles, 1975)

Fundamentals of the DS educational training program

- As a consequence of the teaming up of data scientists and data engineers, observed in the professional data science environments, the DS program offers opportunities for specialization. These specializations relate to distinguished skill and competence sets, and to distinguished toolkits. The students get the possibility to define a personalized training program with focus on specific design and analysis techniques.

- Part of the personalized training program can be used for homologation, part is used to prepare for the individual 12 month project in the second year. Homologation can take maximally 200 hours in the total of the education program.

- If homologation is needed, it is mainly executed on basis of the courses (and certification) offered by the open internet source Coursera, guided by university staff members. See Appendix 4.

- The educational program consists of generalist and expert courses by academic lecturers and by lecturers from the professional field, known for their expertise.

- Blended learning is part of the educational structure, where, for instance in the setting of themes, the students acquire knowledge from open source courses taken from Coursera, eBooks, Stack Overflow, and GitHub. Thus students learn how they can develop themselves further in the field.

- All students should do the generalist courses with a minimum load of 100 hours. Generalist courses are presented more than once a year if necessary. For them, no examination is required.

- The expert part of the program consists of independent modules with a 200 hour load each, where a maximum of 50 hours is devoted to class room lectures or tutorials. Students select a minimum of five modules out of the eight offered. Each module is based on a data science theme or other form of practical assignment to be carried out by a team and started up at the start of the module. Students spend 100 hours on the assignment, the remaining 50 hours are used for feedback sessions with the lecturer both during and at the end of the module, and to skill development typically related to team assessment and peer coaching.

- The educational program is built in such a way that full integration of the technical elements of the program with the soft skills is enabled. Typically, these skills relate to self- and group assessment, communication, team work, project planning and reporting. The program has a coach dedicated to shape this integration.

- Students can enter the program “any time”, in practice typically the first of September and the first of February. With this, a strong learning and social environment is created, where junior students are coached by senior students. Since modules are independent, both junior and senior students can participate in a module team. As seniors take up different roles in such project team, integration of the nontechnical skills is highly enforced.

Data challenge weeks

- Intensive 24/7 one week workshop for trainees/applicants to the DS PDEng program.
- Problems introduced by companies or business institutes
- Integration in the Professional Development Skill training

The data challenge weeks are used to assess possible candidates for the program as part of the application procedure. The assessment also involves a motivation and personality test. The data challenge weeks are organized twice a year for a group of 30 to 40 students, who work in teams of five. Teams are composed of trainees and applicants. Each team has a team leader, who is well prepared for her/his task in separate training sessions. See Appendix 3.

Data science themes

- Training in the ability to solve a Data Science related real world problem
- Applying knowledge related to a specific multidisciplinary area of Data Science.
- Projects are worked out in teams consisting of five students.
- Maximum duration of five weeks with a minimum workload of two days a week
- Teams consist of senior and junior students from both specializations

Theme projects are team assignments concentrating on a coherent field within Data Science. They are used to learn to apply knowledge gained in the course modules offered by the program or obtained from other sources such as self-studies and internet courses.

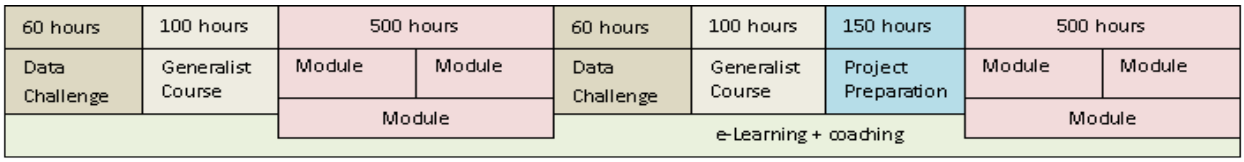
Industrial and business projects

- *Long term project for and within industry and business*
  - Project has multidisciplinary components
  - Project has a duration of maximally 12 months and is executed by one 2nd year trainee
  - Project phasing: definition phase, analyzing phase, conceptualizing phase, modeling phase, implementation phase, implication phase
  - Each phase is concluded with a written report and a presentation
  - Each phase can be started up with a brainstorm session
  - In each phase, the responsible senior (2nd year) trainee can team up with junior (1st year) trainees and (optionally) with (2nd year) master students
  - Dedicated supervision by two staff members, one coaching the project process, one guiding the technical aspects of the project
  - Projects pay 5400 euro per month; execution is (partly) at the company

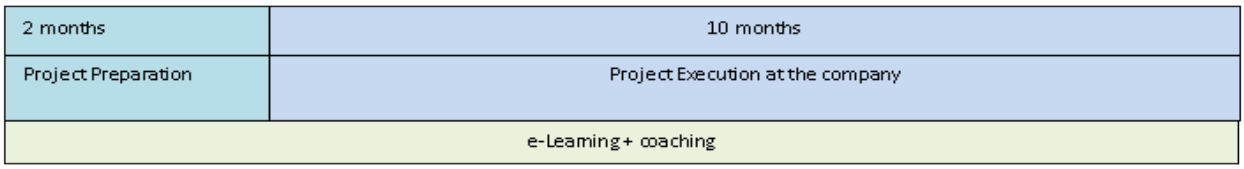
- *Short term projects from industry and business*
  - Working in a multidisciplinary team of four or five trainees on a problem introduced by industry
  - Project duration of maximally four weeks
  - Project planning and time planning related to project phasing
  - Project results reported in project documents and presented in the Data Science Seminar
  - Supervision by one project coach
  - Project pays 3000 euro; execution is at the university

Courses

The generalist courses are presented in a block consisting of six full days within a period of three weeks. The expert courses provided by the program are presented in independent modules within a period of six weeks. The portfolio of courses reflects the T-shape structure of the program. All trainees in the program follow the generalist courses; the expert courses are followed by trainees that want to specialize in certain fields. Part of the courses are taken from e-learning programs available from internet.



Schematic of the first year



Schematic of the second year

Generalist courses

- Focus on the philosophy and generic aspects of Data Science
- Present an overview of the societal role of Data Science
- Explain the contributions of the various disciplines and fields within Data Science
- Teach (mathematical) modeling principles in relationship to use, analysis, and management of data, and the role it plays in tackling problems related to design issues
- Give basic insights in the technical building bricks of Data Science
- Consist of lectures offered by representatives from the professional field
- Are offered twice a year

Generalist Course A – Data Science Landscape

This course offers the students an overview of the data design, management and manipulation tools and processes commonly used by data scientists, and give them insight in how this knowledge is applied to yield measurable business value. Trainees gain an overview of the techniques of data

science, including data analytics, statistical modeling, data engineering, relational databases, SQL and NoSQL, manipulation of data at scale (big data), algorithms for data mining, data quality, remediation and consistency operations.

Generalist Course B - Data Acquisition, Exploration, Mining, Visualization, and Modeling

Only after being effectively collected, structured and cleaned, data can be explored and applicable meaning can be uncovered from data. Trainees learn the basics of processes and tools from machine learning, data mining, data modeling, data visualization and predictive analytics. They get insight in how to collect and gain value from vast amounts of often untapped unstructured data. The course covers the key industry processes for understanding and utilizing data and developing predictive and related models. It presents an overview of related topics including data provenance, privacy, ethics, law and governance.

Expert courses

- Focus on specialist aspects of Data Science related to engineering, entrepreneurship, business, legal and ethics
- Give the opportunity to draw up an individual educational program for each trainee. Each trainee has a personal coach who helps the trainee to select the courses appropriate for her/his specialization
- Expert courses by experts from the program are offered in independent modules
- Part of the expert courses are offered according to the principles of " blended learning, see Appendix 4
- Provide depth through revolving around the data science "life cycle" from data identification to reification down to analytics and feedback loops

Module 1 – Data acquisition, storage, and management

Design of a data management plan and data warehouse - data definition and specification, data quality, data organization, data integration, data integrity, and data equilibrium - , law and ethics, data governance.

Module 2 – Data preparation and exploration

Data quality assessment strategies, data cleansing, data scaling and normalization, data transformation, exploration queries, design of experiments

Module 3 – Data mining and data analysis

Predictive analytics, neural networks, support vector machines, pattern recognition, classification, recognition, clustering

Module 4 – Process mining and process analysis

Process graphics and modeling from event logs, Petri nets, process mining algorithms – discovery, conformance and compliance, enhancement – control flow and resources, group and Interpersonal dynamics in complex organizations

Module 5 – Modeling and simulation

Conceptual modeling, symbolic regression, time series analysis, predictive modeling, design of simulation

Module 6 – Ethics & Law

Data ethics, privacy and human rights, property rights in data (copyright and patents), regulation for data innovation, competition law, data security requirements, data compliance and risk management, legal requirements for data storage, open data, contractual clauses

Module 7 – Data visualization

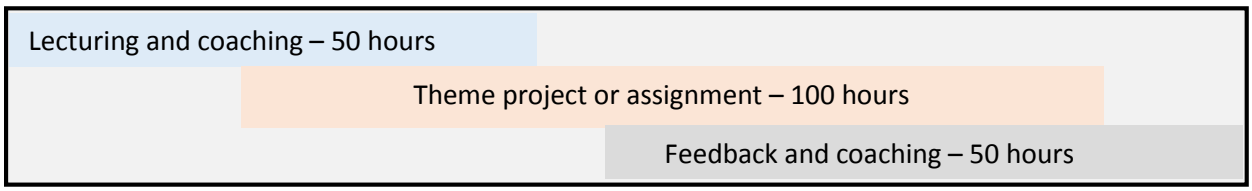
Fundamentals of visual analytics in the practice of communicating with data, design principles of data visualization, human perception, color theory, and effective storytelling, data visualization software and techniques

Module 8 – Business Analytics

Predictive and prescriptive analytics, optimization models and methods, decision making under uncertainty, supply chain analytics, marketing analytics, finance analytics.

Module 9 – Data value chain

Financial information and management systems, information technology policy and strategy, business intelligence, marketing, entrepreneurship, analysis and design of operations to create value, optimization methods for business analytics

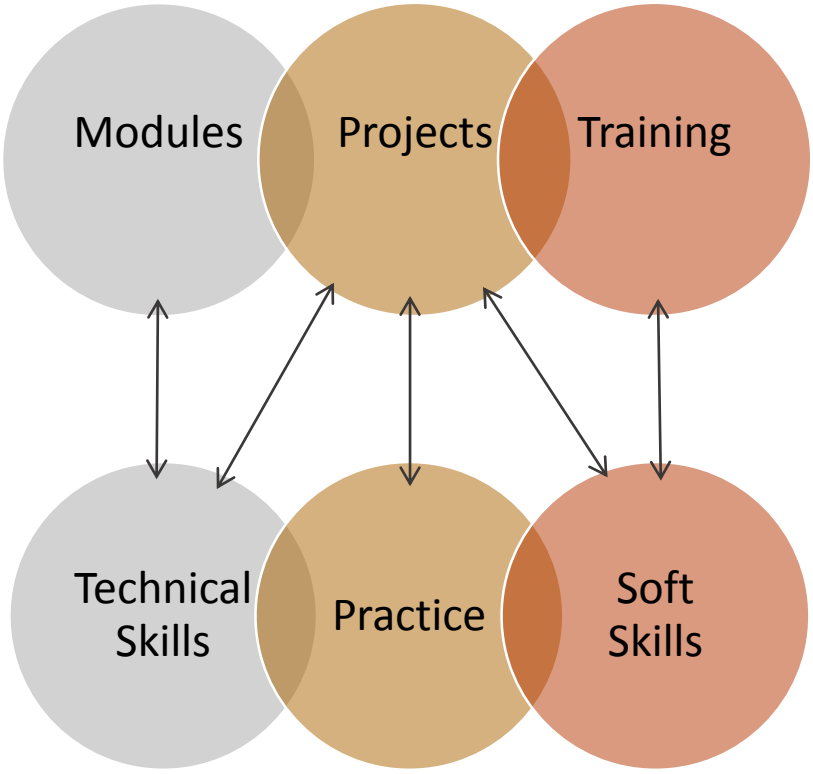


Flow of a module, 200 hours

Soft Skills

Being technically skilled and in possession of adequate technical knowledge is not a guarantee for successfully carrying out Data Science projects in industry, commerce or business. Qualities of character, and social and communicative skills are equally important. In the program, the training and development of social and communicative skills are fully integrated. The coach of the program safeguards these non-technical program elements that are a vital part of the PDEng program DS.

- Selection and admission of candidates is based on assessment both of candidates' technical quality and creativity and of candidates' social and communicative skills. Data challenge weeks are used as the assessment tool.
- At the start of the program, student's learning points and strong points are identified on the basis of the items in a Soft Skill Evaluation Form. Self-reflection and self-assessment is stimulated, by letting the trainees complete their own evaluation form regularly.
- Projects carried out in the frame of the data science practicum offer the natural means to develop conversation skills (discussions with the problem owner), presentation skills (presentation of results takes place in a seminar environment), social skills (projects are carried out by teams of students and are guided by a project supervisor and project coach) and project management skills (the duration of a project is fixed to 3 months with a specified workload; learning to deal with a logbook is one of the training aspects).
- In three-monthly interviews, students' progress in social and communicative skills is evaluated.
- The data challenge weeks and theme projects offer the possibility to focus on team work and team roles. Data challenge weeks are prepared carefully in two or more sessions beforehand by the selected team leaders and evaluated during and after the event.
- During the course of the final project, the non-technical aspects as arrangement of meetings, communication with the problem owner and planning of the project, are discussed individually or in pear coaching sessions.


Modules
Projects
Training
Technical Skills
Practice
Soft Skills

Program management structure

Core Team

The core team consists of:

- The scientific director, the program manager, and the secretary
- Representatives of the main disciplines determining the scientific part of the program.
- The program coach, who takes care of the training on and integration of the soft skills within all the program educational and training elements.

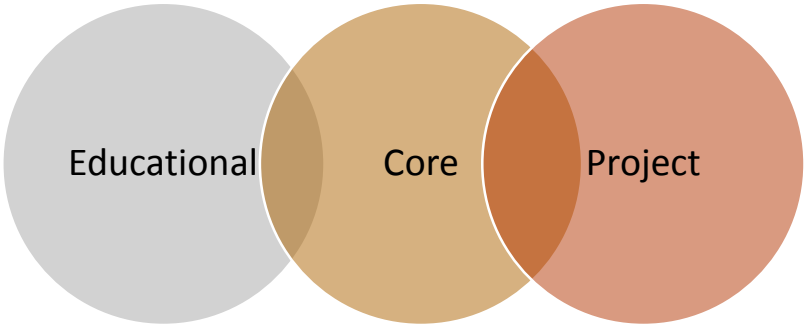
The core team meets regularly, for example, every two months, to discuss the whereabouts of the program and its trainees. The application committee that selects candidates of the program on basis of a positive assessment consists of members of the core team. See Appendix 2.

Educational Team

The educational team consists of the lecturers and coaches who are expert in the various disciplines involved in the program. They take care of the courses specially developed for the program or take care of guiding trainees for that part of the program when content and instruction is delivered by digital or on-line media. The lecturers are from the universities of Tilburg and Eindhoven, or from industry or business. The courses are split up in generalist (broadening) courses and expert courses, presented in modules. Generally more than one lecturer is responsible for a module. Typically, each module contains one or more lectures by an external expert. The educational team meets twice a year to fine-tune their courses. The educational team is represented in the core team by two or three of its members.

Project Team

Acquisition and supervision/coaching of projects and organization of the data challenge weeks are the main tasks of the project team. Members of the project team are carefully selected for their ability to perform the tasks of acquiring, supervising / coaching the projects. The project team is represented in the core team by two or three of its members, who report on the issues related to the acquisition and execution of the projects.



In conclusion

The program strengths

- Cooperation between the universities of Eindhoven and Tilburg within JADS
- Tight connection with the top institute DSC/e of TU/e
- Assessment guaranteeing a thorough selection of candidates through 24/7 data challenge weeks
- Advocated program's T-shape structure reflecting the aim to broaden as well as to specialize, so that graduates of the program are the future Data Science professionals, who oversee the full data cycle its consequences and opportunities in various data domains
- Heterogeneous inflow of students – heterogeneous course offer with the aim to attract the top 25% graduates in Data Science at master level or in related fields as statistics, mathematics, and computer science
- Interdisciplinary, social, learning environment with the emphasis on the students' full responsibility of their own learning process
- Creative, design oriented thinking
- Combined engineering, entrepreneurial, ethical, legal, and business focus by seeking strong cooperation with the ecosystem with respect to various data domains as reflected by the assignments and themes in the modules and invited lectures in the frame of the generalist courses.
- Full integration of professional skills in diverse program elements

*Appendix 1 – **Assessment and evaluation procedures***

- The selection of candidates is based on a thorough assessment in the form of a data challenge week, see Appendix 5. Candidate's motivation, attitude, personality, technical and nontechnical skills are assessed. The assessment is concluded by an application interview.

- Each candidate signs Joint statement of competence assessment and innovation, see Appendix 4.

- Challenge Weeks are workshops and social events also for the students in the program. The program assigns tasks to students to take up in preparation or during the Challenge Week. Performance of the students in the Challenge Week is evaluated by the coach of the program.

- Every six months, an evaluation of the overall performance is arranged. These biannual evaluations are under the supervision of the program manager and the program coach. Part of the evaluation is a discussion of the student's self-assessment of gained knowledge, performance in assignments, and skill progress, see Appendix 5 and 6.

- The course Technical Writing & Editing (TW&E) is integrated within the first year of the program with the emphasis on developing the skills to write technical reports. On basis of the module assignments and data challenge week activities, reports are written by the students. These report writing activities form the basis of the course TW&E and are evaluated as such.

- Performance of students in a module is evaluated by the responsible lecturer of that module. The evaluation is based on the technical quality of the performance, the acquisition and integration of knowledge, and the use of skills.

Example

Students:	M	S	G	E
Evaluation Module #				
What is the quality and extent of knowledge gained?				
What is the overall quality of the assignment?				
How was the team performance in the overall execution of the assignment?				
To what extent have students incorporated knowledge?				
What is the quality of the team presentation(s)?				
What was the quality of the communication? How effective was it?				
To what extent was reflection and peer coaching applied in the process?				
How effective and efficient was the time management?				
What was the quality of team work?				

*Appendix 2 – **Profile and assessment procedure***

Profile

Candidates are creative problem solvers with excellent technical skills. They are team players, who seek interaction with a multidisciplinary environment to tackle problems, they are eager to learn, their attitude shows the responsibility for quality and quantity of knowledge gained. They have a solid background in Mathematics, Statistics, and Computer Science, and strong affinity with Data Science.

1. Candidates possess an academic master degree (MSc) in Mathematics, Statistics, Computer Science, or in a relevant application field of Data Science such as Econometrics and Bio-informatics. The MSc-degree is from an accredited institution comparable to the Eindhoven University of Technology or Tilburg University.

2. Candidates have an academic background that includes:
   - **Mathematics & Statistics –** a least two semesters of mathematics and statistics. Courses on calculus, linear algebra, statistics, optimization theory, and probability theory are required. Courses on topics as operations research, signal and time series analysis, and dynamical systems are recommended. The courses should indicate that the applicant has achieved the mathematical and statistical maturity to be expected of an upper level mathematics / statistics / econometrics graduate.
   - **Computer Science –** at least two semesters of computer science. Courses on programming, algorithms & data structures, databases, data mining/machine learning are required. Courses on topics such as object oriented programming and web development are recommended. The courses should indicate that the applicant has achieved solid knowledge and experience with the computer science aspects that are highly relevant for Data Science.
   - **Data Science –** Candidates show affinity with the field of data engineering or data analytics. They have experience with the application of Data Science technology to real world problems, via projects carried out during their education or their working experience afterwards.

3. Candidates show high **motivation and eagerness** to develop the skills to:
   - Combine generalist thinking and expert thinking in various data domains
   - Formulate opportunities and discover value in data
   - Solve complex problems in a data driven environment.
   - Demonstrate sense for entrepreneurship, business and industrial processes
   - Execute projects in a well-managed, professional way, optimally using resources and obeying constraints
   - Acquire knowledge through a self-directed learning style

4. Candidates have an attitude that shows:
    - **Technical expertise:** desire to become an expert in one or more data domains with a generalist thinking attitude
    - **Curiosity:** desire to discover, distill, and model a problem down to a clear set of concepts and hypotheses that can be tested
    - **Creativity:** desire to look at a problem in different ways and find novel solutions
    - **Communication skills:** desire to communicate effectively with all stakeholders, on all aspects of any data science project that include requirement elicitation, project progress, developed solutions, up to use of data to tell a story.
    - **Social skills:** desire to collaborate with others, with similar and different backgrounds, to tackle challenging problems

In order to select the proper candidates for the PDEng program DS, the management team of the program provides an extensive assessment. This assessment is based on the best practices of the former PDEng program Mathematics for Industry, in which a long history exists from low drop-outs, strong feeling of community, and high learning impact within this assessment.

Assessment procedure

The process of obtaining a position in DS starts by applicants sending a motivation letter and Curriculum Vitae. The first selection is made by the program manager with focus on technical background, specialization, university of applicants' home country, and working experience. The selected candidates are invited for a Data Challenge Week, in which they are assessed on technical and non-technical aspects. The Data Challenge is introduced by a company, business or public institute. Participants work in teams of five.
The Data Challenge Week is a 24/7 social event. After working days, participants are lodged in bungalows. Data Challenge Weeks are organized twice a year. The accepted candidates join the program some months after the Data Challenge Week. The program does not recognize fixed starting dates.
First year students of the DS program take part in the Data Challenge Weeks as well. Whereas this week is an assessment for the applicants, this week offers a learning experience for the DS students. They are challenged to show their acquired technical knowledge, skills on teamwork, leadership and more non-technical issues. After a Data Challenge Week, the feeling of being part of a community increase significantly, which is an important indicator of success.

Assessment criteria

The criteria for assessing applicants in the Data Challenge Weeks are strongly related to the T-profile of the Data Science Professional. Besides the essential technical background applicants are selected on the following distinctive attitudes and skills: creative, self-aware, critical thinking, team and communication skills. For every attitude and skill, assessment criteria are described in more detail.
Essential elements of the assessment:

- Candidates write a report. They write a document on their personal understanding of the research problem and a personal introduction.
- Candidates take part in team activities. They are part of group meetings and technical discussions.
- Candidates give the final presentation. They present results and conclusions of their group to the representatives of the company or business.
- Candidates fill out a personality questionnaire. They fill out the MPT-BS: Multi-Cultural Personality test – Big Six.

Final part of the assessment is the job-interview, a behavioral interview, with a strong focus on technical background, motivation and the assessment criteria. The scientific director, program manager and coach professional development form the interview committee.

*Appendix 3 – **Open source courses***

Inventory of open-source program elements

The DS program will make use of open-sources in its curriculum. For this, the internet can be an oyster. The program will take advantage of Coursera, e-books, Stack Overflow, and GitHub that are all free and open. No doubt, the quality and level of courses has to be constantly monitored. To this end, the following measures will be taken:

- Only courses from high quality providers are selected. In general, Coursera delivers high quality courses, often by top researchers from outstanding universities. The European Data Science Academy (http://edsa-project.eu), in which TU/e participates, is another resource for high quality courses.
- Courses will be judged on content and level by experts in the relevant domains from TiU and TU/e, using a standard assessment form.
- Courses will be evaluated by the students, such that insight in quality, relevance and content of courses from our own target audience is constantly monitored.

Students can use online and open source courses for homologation and during modules. The evaluation of these modules is based on assignments. The program will assess whether the knowledge level of its students is satisfactory. In turn, this provides useful indirect information about the quality of the courses followed.

List of relevant courses:

Generalist courses

1. **Intro to Data Science** UW / Coursera

Topics*:* Python NLP on Twitter API, Distributed Computing Paradigm, MapReduce/Hadoop & Pig Script, SQL/NoSQL, Relational Algebra, Experiment design, Statistics, Graphs, Amazon EC2, Visualization.

2. **Data Science** / Harvard Video Archive & Course

Topics: Data wrangling, data management, exploratory data analysis to generate hypotheses and intuition, prediction based on statistical methods such as regression and classification, communication of results through visualization stories, and summaries.

3. **Data Science with Open Source Tools** (Book $27)

Topics*:* Visualizing Data, Estimation, Models from Scaling Arguments, Arguments from Probability Models, What you Really Need to Know about Classical Statistics, Data Mining, Clustering, PCA, Map/Reduce, Predictive Analytics.
Example Code in: R, Python, Sage, C, Gnu Scientific Library

4. **Data Science as a Profession**

Doing Data Science: Straight Talk from the Frontline O'Reilly / Book $25

5. **Capstone Project**

Capstone Analysis of Your Own Design; Quora's Idea Compendium
Healthcare Twitter Analysis Coursolve & UW Data Science
Analyze your LinkedIn Network Generate & Download Adjacency Matrix

Expert courses

a) Linear Algebra & Programming

Linear Algebra / Levandosky Stanford / Book $10
Linear Programming (Math 407) University of Washington / Course

b) **Statistics**

Statistics I Princeton / Coursera
Stats in a Nutshell Book $29
Think Stats: Probability and Statistics for Programmers Digital &Book $25
Think Bayes Digital & Book $25

c) Differential Equations & Calculus

Differential Equations in Data Science Python Tutorial

d) Problem Solving

Problem-Solving Heuristics "How To Solve It" Polya / Book $10

Get your environment up and running with the Data Science Toolbox

e) Algorithms

Algorithms Design & Analysis I Stanford / Coursera
Algorithm Design, Kleinberg & Tardos Book $125

f) **Distributed Computing Paradigms**

See introduction to Data Science UW / Lectures on MapReduce
Intro to Hadoop and MapReduce UW / Lectures on MapReduce
Hadoop: The Definitive Guide Book $29

g) Databases

Introduction to Databases Stanford / Online Course
SQL School Mode Analytics / Tutorials
SQL Tutorials SQLZOO / Tutorials

h) Data Mining

Mining Massive Data Sets / Stanford Coursera & Digital & Book $58
Mining The Social Web Book $30
Introduction to Information Retrieval / Stanford Digital & Book $56
OSDSM Specialization: Web Scraping & Crawling

i) Machine Learning

Machine Learning Ng Stanford / Coursera
A Course in Machine Learning UMD / Digital Book
The Elements of Statistical Learning / Stanford Digital & Book $80 & Study Group
Machine Learning Caltech / Edx
Programming Collective Intelligence Book $27
Machine Learning for Hackers ipynb / digital book
Intro to scikit-learn, SciPy2013 youtube tutorials

j) Probabilistic Modeling

Probabilistic Programming and Bayesian Methods for Hackers Github / Tutorials
Probabilistic Graphical Modeling Stanford / Coursera

k) Deep Learning (Neural Networks)

Neural Networks Andrej Karpathy / Python Walkthrough
Neural Networks U Toronto / Coursera

l) Social Networks & Graph Analysis

Social and Economic Networks: Models and Analysis / Stanford / Coursera
Social Network Analysis for Startups Book $22

m) Natural Language Processing

From Languages to Information / Stanford CS147 Materials
NLP with Python (NLTK library) Digital, Book $36

n) Analysis

Python for Data Analysis Book $24
Big Data Analysis with Twitter UC Berkeley / Lectures
Exploratory Data Analysis Tukey / Book $81

o) Data Design and Visualization

Envisioning Information Tufte / Book $36
The Visual Display of Quantitative Information Tufte / Book $27
Data Visualization University of Washington / Slides & Resources
Berkeley's Viz Class UC Berkeley / Course Docs
Rice University's Data Viz class Rice University / Slides
D3 Library / Scott Murray Blog / Tutorials
Interactive Data Visualization for the Web / Scott Murray Online Book & Book $26

p) Python

Learn Python the Hard Way Digital & Book $23
Python Class / Google
Think Python Digital & Book $34
Introduction to Computer Science and Programming MIT OpenCourseWare / Lectures
Installing Basic Packages Python, virtualenv, NumPy, SciPy, matplotlib and IPython & Using Python Scientifically
Command Line Install Script for Scientific Python Packages
Pandas Cookbook (data structure library)

q) Data Structures & Analysis Packages

Flexible and powerful data analysis / manipulation library with labeled data structures objects, statistical functions, etc pandas & Tutorials Python for Data Analysis / Book

r) Machine Learning Packages

scikit-learn - Tools for Data Mining & Analysis

s) Networks Packages

networkx - Network Modeling & Viz

t) Statistical Packages

PyMC - Bayesian Inference & Markov Chain Monte Carlo sampling toolkit
Statsmodels - Python module that allows users to explore data, estimate statistical models, and perform statistical tests
PyMVPA – Multivariate Pattern Analysis in Python

u) Natural Language Processing & Understanding

NLTK – Natural Language Toolkit
Gensim -Python library for topic modelling, document indexing and similarity retrieval with large corpora. Target audience is the natural language processing (NLP) and information retrieval (IR) community.

v) Visualization Packages

matplotlib - well-integrated with analysis and data manipulation packages like numpy and pandas
Orange -Open source data visualization and analysis for novice and experts. Data mining through visual programming or Python scripting. Components for machine learning. Add-ons for bioinformatics and text mining

w) Python Data Science Notebooks

Data Science in IPython Notebooks (Linear Regression, Logistic Regression, Random Forests, K-Means Clustering)

Resources

DataTau – The "Hacker News" of Data Science
Metacademy – Search for a concept you want to learn
Coursera – Online university courses
Wolfram Alpha – The smart number and info cruncher
Khan Academy – High quality, free learning videos
Wikipedia – The free encyclopedia
The Signal and The Noise - Nate Silver Pop-Sci Data Analysis $15
Josh Wills - The Life of a Data Scientist / Video
Data Scientist Interviews Metamarkets

Appendix 4 – ***Joint statement of competence assessment and innovation***

Hereby, the PDEng program Data Science (DS), represented by the Program Coach, and the undersigning Trainee certify that with this *Joint statement of competence assessment and innovation* details of the competences are set out that Data Science PDEng trainees would be expected to innovate during their training.
We state that the competences may be present on commencement, explicitly taught, or developed during the course of the program. It is expected that different mechanisms will be used to support learning as appropriate, including self-direction, coach support, workshops, brainstorming sessions, modules and data challenge weeks.
The management of the PDEng program DS re-emphasizes its belief that training in competences is a key element in the development of the trainee, and that trainees are expected to make a substantial, original contribution to self-knowledge, normally leading to self-assessment. The development of wider employment-related competences should not detract from that core objective.

The statement gives the program's common view of the competences that a trainee in the PDEng program DS should acquire during the stay in the program in relation to:

- General rules of conduct
- Self-reflection
- Conversation skills
- Presentation skills
- Social skills and team work effectiveness
- Self-directed learning attitude
- Project management skills
- Career management

Signature Trainee: Signature Coach:

's Hertogenbosch, d.d.

See Enclosure.

Enclosure: Joint Statement of Competence Assessment and Innovation

General Rules of Conduct:

Show actively one's commitment in respect to the Postmasters Program.
Demonstrate awareness of issues relating to the rights of other data scientists, of the data science seminars and modules, and of others who may be affected by the program activities, e.g. confidentiality, ethical issues, integrity, attribution, copyright, malpractice, ownership of data and the requirements of the Data Protection Act.
Understand the processes for funding and evaluation of Personal Competence Innovation.
Constructively defend outcomes at seminars and in industry.
Contribute to promoting the public understanding of one's data science field.

Self-reflection - to be able to:

Formulate personal learning points and strong points
Be conscious of own actions and distinguishing the effect they have on others
Be conscious of own opinion and being able to consciously react on the basis of this opinion
Be conscious of one's own (none) verbal communication and being able to react on communication of others

Conversation Skills – to be able to:

Formulate clear goals preceding a conversation
Listen, to pose relevant questions and to summarize during a conversation
Negotiate and to persuade
Actively communicate with the problem owner

Presentation Skills – to be able to:

Construct coherent arguments and articulate ideas clearly to a range of audiences, formally and informally through a variety of techniques.
Give an informative and convincing presentation, well-tuned to the audience
Present with a positive attitude
Give a well-organized presentation

Social Skills and Team Work Effectiveness – to be able to:

Take initiative
Participate actively and assertively in a meeting
Come to decisions
Develop and maintain co-operative networks and working relationships with supervisors, industrial co-workers and peers, within the institution and the wider Research & Development industry.
Understand one's behavior and impact on others, when working in and contributing to the success of formal and informal teams.
Listen, give and receive feedback and respond perceptively to others.
Manage conflicts
Effectively support the learning of others when involved in mentoring activities.

Learning Styles – to be able to:

Demonstrate willingness and ability to learn and acquire knowledge.
Be creative, innovative and original in one's approach to data science activities.
Demonstrate flexibility and open-mindedness.
Demonstrate self-awareness and the ability to identify own training needs.
Demonstrate self-discipline, motivation and thoroughness.
Recognize boundaries and draw upon/use sources of support as appropriate.
Show initiative, work independently and be self-reliant.

Project management – to be able to:

Apply effective project management through the setting of goals, intermediate milestones and prioritization of activities.
Show leadership
Summarize the essence of a meeting
Summarize, document, report and reflect on progress.

Career Management – to be able to:

Appreciate the need for and show commitment to continued professional development.
Take ownership for and manage one's career progression, set realistic and achievable career goals, and identify and develop ways to improve employability.
Demonstrate an insight into the transferable nature of human resource skills to other work environments and the range of career opportunities within and outside academia and or industry.
Present one's competences, personal attributes and experiences through effective CV's applications and interviews.

Appendix 5 – Four Monthly evaluation

Name student:
Start Date:
Date of Evaluation:

EDUCATIONAL PROGRAM

Status: [please indicate the courses completed and running, including lecturer, end date, assignment and partner if applicable]

Courses & Modules	Lecturer	End Date	Topics	Assignment Status	Partners

Previous agreements:
New agreements:

TECHNICAL WRITING & EDITING
Status: [please indicate the assignments of the course completed, including date and running]

	Assignment	Finished Date
1		
2		
3		

Previous agreements:
New agreements:

PERSONAL DEVELOPMENT
Status: [please indicate the trainings completed]

	Data Challenge Week & Coach Meeting	Date
1		
2		
3		
4		

Previous agreements:
New agreements:

EXTRA-CURRICULAR ACTIVITIES
Activity: [please indicate the activity and date/period of completion]

	Activity	Date
1		
2		
3		

Other Points of Discussion
[please indicate points of discussion about the program and your development]

Appendix 6 – Self-assessment form - social and communicative skills

Self-reflection	M	S	G	E
• Conscious of personal learning points and strong points				
• Conscious of own actions and distinguishing the effect they have on others				
• Conscious of and being able to consciously react on the basis of own opinion				
• Conscious of own communication and able to react on communication of others				

Conversation skills				
• Able to formulate clear goals preceding a conversation				
• Able to listen, to pose relevant questions and to summarize during a conversation				
• Able to negotiate and to persuade				
• Able to actively communicate with the problem owner				

Presentation skills				
• Able to give an informative and convincing presentation, well-tuned to the audience				
• Able to present with a positive attitude				
• Able to give a well-organized presentation				

Social skills				
• Showing initiative				
• Able to participate actively and assertively in a meeting				
• Able to motivate and stimulate team members				
• Able to deal with disagreements in a team				
• Able to come to decisions				
• Able to give and receive feedback				

Project management skills				
• Showing leadership				
• Able to summarize the essence of a project meeting				
• Able to carry out project and time management				
• Able to keep a logbook				