Addressing data acumen with an introduction to data science

Christin Schmidt

Hochschule für Technik und Wirtschaft (HTW) Berlin, Germany

christin.schmidt@htw-berlin.de

Abstract

This paper describes an approach to establishing and teaching an introductory module to data science within the undergraduate bachelor degree programme Applied Computer Science at Hochschule für Technik und Wirtschaft (HTW) Berlin, University of Applied Sciences. The module is part of an elective study program data science. In consequence, the introductory module has to serve subsequent modules while also having to address general interdisciplinary foundations in mathematics, statistics, computation, ethics and practical domain-specific applications. Thus, a syllabus, a module structure and a teaching approach combining theoretical foundations, practical exercises and supplementary practical student projects aligned to a proposed workflow are presented. Additionally, the module is assessed with regard to student evaluations and a taxonomy of data acumen adapted from the United States National Academies of Sciences, Engineering, and Medicine as well as the EDISON Data Science Framework.

I. Introduction

As domains in science and industries are increasingly data driven, a constantly growing demand for skilled data scientists has emerged over the past years. Many universities have adapted to this by either having integrated or planning to integrate data science content into their curricula, e.g. through offering related modules, study or full degree programmes at different levels. Likewise, a need for an introductory course to data science within the curriculum of the undergraduate bachelor degree program of Applied Computer Science at Hochschule für Technik und Wirtschaft Berlin was identified in 2015. In the following, the module design is described in view of the conditions and module goals. The proposed syllabus and teaching approach are presented. The module is evaluated by presenting student evaluations and pursuing the question how the established module contributes to proposed key concepts and skills associated with acumen in data science education.

II. Module Design

The module is part of the curriculum in Applied Computer Science (Bachelor) with the elective study program *data science*, which in total contains five modules in different semesters as follows:

- *Introduction to Data Science* (3rd sem.),
- *Machine Learning* (4th sem.),
- *Development of Data Science Applications* (4th sem.),
- *Data and Text mining* (5th sem.), and
- *Selected chapters of Data Science* (5th sem.).

The module consists of two units, a seminaristic lecture (SL) and a (lab) exercise (E). The module units run weekly for 90 min. each, together amounting to five credit points (ECTS). The planned capacity of the module amounts to 22 students enrolled in total. The module is open for enrollment to students with different skill levels from other study programmes and departments.

i. Goals of the module

In general, the module aims at building foundations and student's abilities to be able to work as a data scientist and build on their knowledge.

The module's goals are to provide basic theoretical knowledge and practical skills to enable students to work in data science projects and to pursue advanced aspects of data science in subsequent modules. Hence, the module has to address mathematical, statistical and computational foundations as well as providing an option for gaining practical experience by participating in a domain-specific project. In addition, the module should provide guidance with regards to relevant project workflow stages as well as aspects of teamwork, ethics, law and scientific methodology.

In order to achieve the goals, a course structure, syllabus and teaching approach considering the circumstances explained above had to be developed. In the following, the approach to meeting the underlying goals is presented and evaluated.

ii. Syllabus

Given the underlying conditions and goals, a syllabus has been developed for the module. Topics and structure of the seminaristic lectures and exercises are summarized in Table 1.

The contents have been gradually developed in view of the demand of the lecturers of subsequent modules within the elective study program (cp. section II) and the feedback from the student evaluations (cp. subsection ii).

Table 1 shows 20 units in total (as the minimum content to be covered by the module) of which two units are presented weekly (90 mins. for each SL and E). The unit topics of the SL are picked up in the lab exercise.

The exercises build upon the correspondent lecture units and contain two parts: a retrospective part of the lecture and lab exercises. The first retrospective part repeats concepts from the lecture. The second part consists of exercises with a focus on applying concepts of the seminaristic lecture via both practical programming in R and transferring methods to the student's specific projects.

The timeline is offering flexibility to stretch specific topics according to the progress in classroom, e.g. the unit „Statistical Foundations [IV]" can be stretched over two weeks or more if necessary.

The remaining units towards the end of each summer and winter semester contain presentations and discussions of the student projects as well as potential slots for excursions, guest lectures or selected chapters of data science.

iii. Teaching Material

Accompanying the module teaching material has been created in German:

- Lecture notes (instructors),
- Lecture notes (students),
- Lab notes (instructors) supplemented by Jupyter/IPython notebooks, and
- Lab notes (students) supplemented by Jupyter/IPython notebooks.

The notes are made accessible during the semester via moodle provided by HTW Berlin[1].

iv. Course examination

In order to be examined, students have to prepare a written project documentation (50% of final grade) and present their project at the end of the semester (50% of final grade).

Grades are examined individually based on the grading scheme of the university according to criteria that are explained in the following.

v. Grading criteria

The grading criteria are communicated to the students as they reflect the structure of the presentation as well as the written report.

Grading is executed in accordance with the grading scheme of HTW Berlin.

[1] Available online: https://moodle.htw-berlin.de [last accessed: 8. Oct. 2019]

Table 1: *Syllabus*

Unit	Contents (SL)	Contents (E)
Introduction	Terminology: data, data scientist, examples of data science applications, skills and competencies in data science, the working process of a data scientist, limitations of data analyses, faulty data, the nature of statistics	Part 1: Lecture retrospective; Part 2: Lab exercises: - Terminology, - Historical aspects of data science, - Student team project: Phase I (Discovery): team forming and topic exploration.
Data analysis: A workflow perspective	Process (Discovery, Acquisition, Preparation, Exploration/Model Planning, Model Building, Interpretation, Publication, Operationalization), activities, goals and milestones	Part 1: Lecture retrospective; Part 2: Lab exercises: - How to formulate a good research question, - Data acquisition.
Ethics & data privacy	Terminology: ethics & moral responsibility, influencing factors for ethical judgement, assessment and reflexion, ethical influence to data science, lying with statistics, codices, legal frame, personal data and personal rights, key principles of the General Data Protection Regulation	Part 1: Lecture retrospective; Part 2: Lab exercises: - Case study, - Ethical aspects of the student team project.
Workshop: student project	Part 1: Conditions of the student team project; Part 2: Work on student team projects	Part 1: Work on student team projects; Part 2: Presentation/discussion of results (short project profiles)
Statistical foundations [I]	Terminology, scale levels, data exploration and descriptive analysis, measures of central tendency, measures of variation, distribution patterns, data visualization	Part 1: Lecture retrospective; Part 2: Lab exercises: - Loading and modifying data (R), - Measures of central tendency and variation, distribution patterns (R), - Scale levels.
Statistical foundations [II] Association and correlation	Frequencies and contingency tables, (odds) ratio, observed and expected frequencies, association and measures, correlation and measures, interpreting measures, limitations of measures	Part 1: Lecture retrospective; Part 2: Lab exercises: - χ^2 analysis and significance levels with simple vectors (R), - Correlation analysis (R).
Statistical foundations [III] Regression models	Regression models: types, domains used, aspects of formalising, estimating parameters, assessing model quality, overfit vs. underfit	Part 1: Lecture retrospective; Part 2: Lab exercises: - Linear regression (R), - Work on student team projects.
Statistical foundations [IV] Probability, inference and test theory	Statistical inference, chance, probability, Bayes' theorem, density functions and curves, distributions, test theory, falsification, formulating and testing hypotheses, bias & error, significance level	Part 1: Lecture retrospective; Part 2: Lab exercises: - Hypotheses & t-test (R), - Work on student team projects.
Machine Learning [I] Supervised Learning	Terminology: learning types (supervised, unsupervised), methods and applications of supervised learning, prerequisites, learning algorithms, classification problems, classificator, discriminant and version space	Part 1: Lecture retrospective; Part 2: Lab exercises: - Decision trees: splitting data, learning phase, plotting, predicting (R, libraries: rpart, rpart.plot), - Work on student team projects.
Machine Learning [II] Unsupervised Learning	Methods and applications of unsupervised learning, learning algorithms, distance and distance metrics, dimensionality reduction, clustering	Part 1: Lecture retrospective; Part 2: Lab exercises: - K-means clustering (R), - Work on student team projects.

Adequacy and precision of the documented and presented approaches are assessed in each of the dimensions as follows:

- Problem statement/research question(s),
- Theoretical foundations/literature review,
- Methodology: planned approach of the analysis,
- Executed approach of the analysis,
- Findings,
- Conclusions,
- Limitations,
- Future Work, and
- Sources.

III. Data science workflow

Given the conditions and goals explained above, the module has to integrate general foundations in combination with offering experiences with a feasible domain-specific practical project. In this context, the goal of guiding students along the process of problem-solving and different phases of a data science project arose. Yet, when starting the module there was no general framework available in the data science discipline, which could be transferred and consistently applied into teaching practice.

For example, CRISP-DM [10, pp. 31ff.], a business-oriented process model for data mining, introduces the phases: 1. Business Understanding, 2. Data Understanding, 3. Data Preparation, 4. Modeling, 5. Evaluation, 6. Deployment.

Janssens [7, pp. 2ff.] distinguishes between: 1. Obtaining data, 2. Scrubbing data, 3. Exploring, 4. Modeling, and 5. Interpreting data.

Dietrich et al. [5, pp. 29ff.] introduce further phases as follows: 1. Discovery, 2. Data Preparation, 3. Model Planning, 4. Model Building, 5. Communicate Results, 6. Operationalize.

O'Neil and Shutt [9, pp. 41ff.] differentiate between: 1. Raw data collection, 2. Data Processing, 3. Data Cleaning, 4. Exploratory Data Analysis, 5. Machine Learning Algorithms/Statistical Models, 6. Communication, Visualization, Reporting, 7. Building Data Products.

Bell [1, pp. 17ff.] describes 1. Acquisition, 2. Prepare, 3. Process, 4. Reporting.

The teaching context demands slightly other and/or more phases than those presented above.

As an example, students do not start with a business understanding as recommended in CRISP-DM. Instead, they have to explore and find a research question and set up a methodological plan before data is obtained. This aspect is reflected in the phase „Discovery" in Dietrich et al. [5, pp. 29ff.]. It is not reflected by Janssens [7, pp. 2ff.], O'Neil and Shutt [9, pp. 41ff.] and Bell [1, pp. 17ff.].

Interpreting data is described as a single phase by Janssens [7, pp. 2ff.] and appears to be an important competence throughout the workflow. Reviewing and interpreting both the process and the process results should also be addressed in teaching data science from the author's point of view. None of the approaches above addresses this.

Thus, missing a suitable workflow approach, that could easily be transferred to the classroom, an own workflow was tailored and evolved as a blend of the approaches above (Figure 1). This was done with the aim of giving student's a structure guiding them through the working process of a data science project[2].

i. Exploration/Discovery

The first phase embraces the discovery and the formulation of a research question as well as the establishment of goals of a potential analysis. This also involves a first exploration and review of relevant existing work in order to assess the ability of the own research project to build upon existing knowledge, models, hypotheses, methodologies and data.

In addition, it has to be described, which data, objects and carriers of characteristics are in focus and available.

Furthermore, a first plan of how to answer the underlying research question has to be es-

[2]The approach to a workflow presented here by no means claims to be generic or transferrable to various other domains and contexts. Future research could address this identified lack in theory.

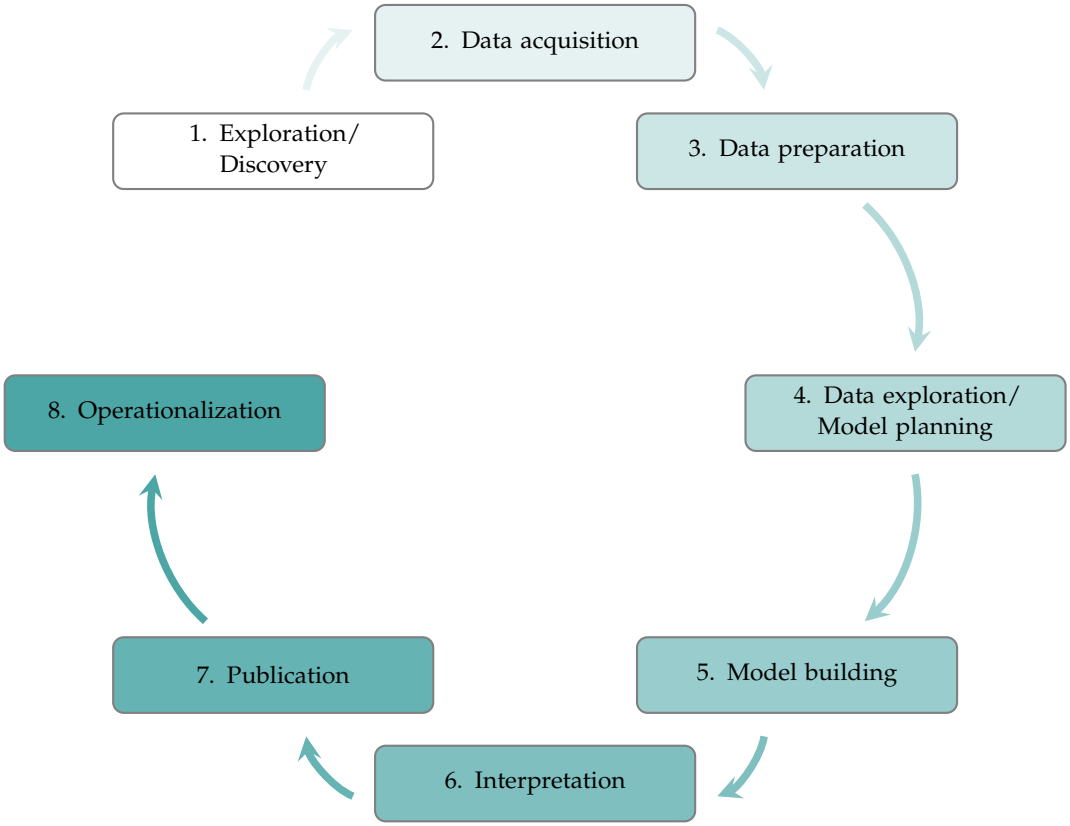

Figure 1: *Data Science Workflow*

tablished in view of contextual factors. Context comprises many different aspects, which may or must be weighted differently, depending on the specific nature and domain of the project, for example:

- Resource context: discipline(s), theory, infrastructure, technologies, data sources, methodologies, methods, skills,
- Social context of/interaction context with stakeholders of the data science project,
- Legal context, and
- Ethical context.

The phase should aim at narrowing down the research problem with the goal to achieve the following milestones and artifacts:

- A formulated research question: this can be either explorative in order to find hypotheses or explanative by testing and/or applying existing hypotheses or models respectively.
- A rough structure and working plan of the project according to the lifecycle describing the underlying methodology in order to answer the research question.
- Identification, initial evaluation and establishment of an inventory of needed data in view of availability as well as contextual conditions.

ii. Data acquisition

The phase data acquisition is executed according to the prerequisites and the defined methodology established and described in the previous phase. It aims at acquiring and storing the adequate amount and quality of raw data for subsequent phases.

iii. Data preparation

The preparation of data comprises the porting of collected data into a previously defined working space (sandbox) accompanied by a deeper analysis of the quantity and quality of the obtained data. This may involve further activities of extraction and modification of data, e.g.:

- Normalizing: converting data types, structures and formats;
- Aggregation/consolidation of data (sets),
- Splitting of data (sets), and
- Data cleaning: handling of outliers, missing data, flawed data, inconsistent data.

The phase should provide the data for subsequent activities in alignment to the specific context and the underlying methodology.

iv. Data exploration/Model planning

Independent from the nature of the research question, the phase data exploration/model planning comprises activities with the aim to obtain an abstraction of the reality that is contained in the data.

First steps in getting to know the data relevant for subsequent analysis comprise the exploration of the distribution of the data. Supplementary, visualizations, calculation of measures of tendency and variation in order to get an overview of the data shall be realized. Depending on the nature of the research question the goal could be to explore first characteristics within the data, eventually ending with working hypotheses for future work. If so, the workflow proceeds with phase 6 (Interpretation).

In case there is a model/hypothesis based on either the previous phases or as an interim result of the exploration, the workflow proceeds with planning towards a specific model as a prerequisite for the next phase. For example, this could involve a selection of variables for further analysis and tests towards association, correlation or setting up a first regression hypotheses without fitted parameters.

In case a specific model is associated with the research question and underlying methodology, the model planning phase is reduced or even omitted in favor of proceeding directly to the subsequent phase.

v. Model building

This phase addresses the establishment, test and potential refinement of a model or (a set of) hypotheses with the goal to answer the research question. For example, this could comprise comparing distributions in samples, predicting adequate values with regression models or applications of machine learning, such as supervised classifications with artificial neural networks.

Thus, various further activities might have to be considered, e.g. training of a model to estimate parameters, fitting of estimated parameters, testing, refinement, application and evaluation of a model.

vi. Interpretation

The phase interpretation is addressing the postulation to draw conclusions from both the process and the results discovered through the process in view of the underlying research question.

Process and process results should be reviewed and reflected according to the specific phase in order to be able to spot the lessons learned. This should address both aspects on how the project contributed to answering the research questions and limitations of the work or the results.

vii. Publication

The phase publication comprises the communication of results. Given a solid documentation throughout the lifecycle, this could embrace the publication of the whole project or partial project aspects. In view of the stakeholders defined in the first discovery phase (cp. subsection i), target groups as well as channels of publication may vary.

viii. Operationalization

The phase operationalization addresses subsequent activities to the realized data science project as a result from the findings of the phase interpretation (cp. subsection vi).

Future work might comprise:

- Development and/or refinement of model(s),
- Application of different methodology, data or workflow to answer the same research question,
- Application of the same methodology, data or workflow plan to answer a different research question,
- Application of the model to other contexts,
- Pursuit of further development of theory, and
- Pursuit of further practical development, e.g. data scientific applications, products or product components.

ix. Superior skills and activities

The iteration through a data science project involves various other general and context-specific skills that have to be bundled into supporting activities and processes. These might be applied to either one, some or all process phases of the data science workflow (cp. Figure 1).

Besides suitable skills in mathematics, statistics and computation it seems also reasonable to derive and glean supporting processes from the domain-specific context. As an example, ethical and/or legal aspects might influence the approach to data collection, data preparation and data management as a whole.

Industrial project settings might postulate further management activities (project, team, software engineering) or the application of specific (process) models along the workflow.

IV. Module Assessment

i. Data acumen

Initially, the module design was not able to draw upon a general framework or concept for the education of future data scientists. Meanwhile, some approaches are available. The United States National Academies of Sciences, Engineering, and Medicine [8, pp. 22-33] point out that a critical task in the education of future data scientists is to instill *data acumen*, which includes the following key concepts:

- Mathematical foundations,
- Computational foundations,
- Statistical foundations,
- Data management and curation,
- Data description and visualization,
- Data modeling and assessment,
- Workflow and reproducibility,
- Communication and teamwork,
- Domain-specific considerations, and
- Ethical problem solving.

The EDISON Data Science Framework extends this perspective. It proposes two further areas of competency characterized as highly demanded and specific to data science: 1. Preservation (in addition to data management and curation) and 2. Scientific or Research Methods [3, pp. 622f.].

Although the approaches focus on data science programs rather than single modules, they seem able to serve as a basis for an assessment of how the module described here matches the proposed key concepts and skills associated with data acumen.

In the following, the key concepts are sequentially described and transferred onto the module in order to assess how it contributes to the above mentioned dimensions associated with data acumen.

i.1 Mathematical foundations

It is proposed that data scientists need to know how to test hypotheses, how to determine and assess their data science models, and how to make corrections that lead to scientific discovery. Several basic mathematical concepts and skills are highlighted as important for students in data science programs. These comprise:

- Set theory and basic logic;
- Multivariate thinking via functions and graphical displays,

- Basic probability theory and randomness,
- Matrices and basic linear algebra,
- Networks and graph theory, and
- Optimization.

Additionally, advanced mathematical concepts that should be included in study programs are mentioned as follows:

- Partial derivatives (to understand interactions in a model),
- Advanced linear algebra (i.e. properties of matrices, eigenvalues, decompositions),
- *O* notation and analysis of algorithms, and
- Numerical methods (e.g. approximation and interpolation).

The curriculum of the study course of Applied Computer Science addresses many (advanced) mathematical skills at the bachelor's and master's level.

Supplementary, the module described here contributes to addressing mathematical foundations by:

- Integration of mathematical foundations throughout the module according to syllabus of course units (cp. Table 1), e.g. test theory, (multivariate) regression functions and basic probability theory and randomness, and
- Practical application of mathematical foundations in combination with statistical and computational foundations in both lab exercises and student projects.

Yet, graph theory and network analysis have not been addressed by the module so far.

i.2 Computational foundations

Data science students should develop an ability for algorithmic thinking and abstraction by addressing:

- Basic abstractions,
- Algorithmic thinking,
- Programming concepts,
- Data structures, and
- Simulations.

A wide range of computational skills is addressed by the curriculum of the study course. In addition to skills established in other modules, the module described here contains theoretical and practical computation foundations as shown in the syllabus of the lectures and lab exercises (cp. Table 1).

Students are free to use further computational knowledge (e.g. paradigms, languages, frameworks and technologies) for their projects in addition to those presented in the module.

i.3 Statistical foundations

Important statistical foundations [8, p. 25] embrace:

- Variability, uncertainty, sampling error, and inference;
- Multivariate thinking,
- Nonsampling error, design, experiments (e.g., A/B testing), biases, confounding, and causal inference;
- Exploratory data analysis,
- Statistical modeling and model assessment, and
- Simulations and experiments.

Addressing statistical skills is one of the key targets of the study program. In addition to skills established in other modules, the module described here addresses all of the above mentioned basic statistical foundations.

i.4 Data management and curation

Key concepts and skills associated with data management and curation are suggested as follows:

- Data provenance,
- Data preparation (data cleansing and transformation),
- Data management (of various data types),
- Record retention policies,
- Data subject privacy,
- Missing and conflicting data, and
- Modern databases.

The module does not focus on data preservation as postulated by the EDISON Data Science Framework [3, pp. 622f.]. While the field of databases is covered in more detail in other modules in the full degree study course, all remaining concepts are addressed theoretically

and practically in the module course units (cp. Table 1) by both seminaristic lectures, lab examinations and student projects.

i.5 Data description and visualization

Foundations in traditional descriptive statistics and graphics should comprise:

- Data consistency checking,
- Exploratory data analysis,
- Grammar of graphics,
- Static and dynamic visualizations, and
- Dashboards.

Addressing skills in data description and visualization is one of the key targets of the study program.

In addition to skills established in other modules, the module described here seems to address the above mentioned basic foundations theoretically and practically in the module course units (cp. Table 1) by both seminaristic lectures, lab examinations and student projects, while excluding dynamic visualizations and dashboards.

i.6 Data modeling and assessment

Key concepts related to data modeling and assessment should comprise:

- Machine learning,
- Multivariate modeling and supervised learning,
- Dimensionality reduction techniques and unsupervised learning,
- Deep learning,
- Model assessment and sensitivity analysis, and
- Model interpretation.

Addressing skills in data modeling and assessment is one of the key targets of the study program. In addition to skills established in other modules, the module described here seems to address the above mentioned basic foundations theoretically and practically in the module course units (cp. Table 1) by both seminaristic lectures, lab examinations and student projects.

As deep learning is a focus of subsequent modules within the study program, it is excluded in the presented module.

i.7 Workflow and reproducibility

It is suggested that an *„effective data science workflow involves formulating good questions, considering whether available data are appropriate for addressing a problem, choosing from a set of different tools, undertaking analyses in a reproducible manner, assessing analytic methods, drawing appropriate conclusions, and communicating results“* [8, p. 20].

In order to do so, students should be presented a unified and integrated approach in their first courses, which remains consistent in subsequent courses.

In this context, key concepts that should be addressed comprise [8, pp. 27f.]:

- Workflows and workflow systems,
- Documentation and code standards,
- Source code (version) control systems,
- Reproducible analysis, and
- Collaboration.

The module seems to address workflow and reproducibility in the syllabus by building theoretically on the proposed workflow as described above (cp. section III).

Practical workflow, documentation, reproducibility and collaboration skills are also addressed by the student projects. Students are free to use source code (version) control systems, that they have worked with before in other modules with a focus on programming and software engineering/development.

i.8 Communication and teamwork

Key concepts related to communication and teamwork are suggested to include:

- Ability to understand client needs,
- Clear and comprehensive reporting,
- Conflict resolution skills,
- Well-structured technical writing, and
- Presentation skills.

The module seems to address the above mentioned key concepts to a great extent by integrating interactive parts into lectures and labs as well as student projects.

Yet, the ability to understand client needs is not addressed by the module, since the topics and domains of the student projects are chosen by the student teams, not by specific clients.

i.9 Domain-specific considerations

According to the United States National Academies of Sciences, Engineering, and Medicine an *„[e]ffective application of data science to a domain requires knowledge of that domain. Grounding data science instruction in substantive contextual examples (which will require the development of judgment and background in those areas) will help ensure that data scientists develop the capacity to pose and answer questions with data“* [8, p. 29].

Addressing domain-specific considerations is one of the key targets of both the full degree course of Applied Computer Science and the study program data science.

In addition to skills established in other modules, the module described here addresses the above mentioned foundations by integrating examples from different domains in the syllabus (cp. Table 1) as well as student projects.

Table 2 provides an overview of chosen topics, which imply various domains that had to be considered by the students in their data science projects.

i.10 Ethical problem solving

Ethics and ethical problem solving should be addressed throughout the data science education as follows [8, pp. 30f.]:

- Ethical precepts for data science and codes of conduct,
- Privacy and confidentiality,
- Responsible conduct of research,
- Ability to identify „junk“ science, and
- Ability to detect algorithmic bias.

The module described here seems to address the above mentioned basic foundations theoretically and practically in the module course units (cp. Table 1) by both seminaristic lectures, lab examinations and the student projects.

Students are introduced to general concepts of research and science, e.g. scientific fraud occurring either through data falsification, data massaging or data fabrication [6, pp. 132ff.].

Whilst the community lacks a consistent code of conduct, first approaches to common consistent principles and regulations shall be presented here.

Thus, the Government of the United Kingdom (Department for Digital, Culture, Media & Sport) [4] points out seven principles in context of a Data Ethics Framework as follows:

- Start with clear user need and public benefit,
- Be aware of relevant legislation and codes of practice,
- Use data that is proportionate to the user need,
- Understand the limitations of the data,
- Ensure robust practices and work within your skillset,
- Make your work transparent and be accountable, and
- Embed data use responsibly.

Supplementary, the Regulation (EU) 2016/679 of the European Parliament and of the Council introduced a General Data Protection Regulation (GDPR), that has to be applied since May 2018 within the European Union.

Table 2: *Topics of student team projects (excerpt)*

Semester	Topics
Summer 2016	- Analysis of global music trends on Twitter - Methods of lexical diversity with practical examples - Sentiment analysis of comments on Reddit - Population density and diseases in Germany
Winter 2016/2017	- The narrative of German news - Development of rents in Berlin - Development of life quality in the EU - Exploring the cheapest online retailer in comparison to Amazon - Hate speech on Twitter
Summer 2017	- Comparison of online-journalism portals - Metasearch for cooking recipes - Minimum wage and purchasing power in Germany - Survey on study start for first semester students - Exploring climate change
Winter 2017/2018	- Musical Lyrics - Classical TV vs. Online Streaming - Semantic analysis of comments on Twitter - Comparison of the results of the German elections 2013-2017 - Exploration of publicly available APIs to gather and analyse data
Summer 2018	- Classification of midi-data in classical music - Tweets and their influence on the stock exchange - Analysis of menus in the university dining hall in view of allergical aspects - Fear on Twitter

Article 5 (GDPR) [2] addresses further principles in processing personal data:

- Lawfulness, fairness and transparency,
- Purpose limitation,
- Data minimisation,
- Accuracy,
- Storage limitation,
- Integrity and confidentiality, and
- Accountability.

i.11 Scientific research methods

. According to the EDISON Data Science Framework presented by Demchenko et al. [3, p. 623], knowledge of scientific research methods and techniques are additional competencies to be addressed in data science training and education.

The module described here seems to address this by introducing students to theoretical aspects of scientific research methods (cp. Table 1), and by practically guiding students through the phases of a real scientific data science project according to the workflow presented in this paper (cp. section III).

ii. Student evaluation and enrollment

Every two years, modules are evaluated by students online via a quantitative questionnaire[3].

The used questions and scales include the measurement of a degree of satisfaction[4] with the learning progress in the specific course unit (seminaristic lecture (SL) and (lab) exercise (E)).

Overall, the evaluation results with N respondents are presented in Table 3.

The results show that the respondents seem to be rather to very satisfied with their perceived learning progress.

[3]The questionnaires of the student evaluation are available online: https://www.htw-berlin.de/en/organisational-units/central-offices/university-development-quality-management/evaluation/ [last accessed: 8. Oct. 2019].

[4]The underlying scale is: 1. *very satisfied*, 2. *rather satisfied*, 3. *partly satisfied/dissatisfied*, 4. *rather dissatisfied*, 5. *very dissatisfied*.

Table 3: *Results of student evaluations*

Semester	**Unit**	*N*	**Satisfaction with learning progress in unit**
Summer 2016	SL	11	1.5
	E	n.a.	n.a.
Winter 2017/2018	SL	9	1.6
	E	7	1.4

In addition to the quantitative score, the questionnaires contained open questions, which led to adjustments of the module as follows:

- Summer 2016:
  - Preparation of Jupyter/IPython notebooks with more R-Tutorials, examples and exercises for individual study (cp. Table 1), and
  - Extension of the module syllabus with unit „Ethics & Data Privacy“.
- Winter 2017/2018: Due to a high demand the standard capacity of the course was doubled to 44, which resulted in a larger audience capacity in the lecture and two lab exercise groups.

Although the results seem to show a positive tendency with regard to satisfaction, limitations of the student evaluation embrace the low response rate and the time interval, which do not enable a continuous and representative insight into the module from a student's perspective.

V. Conclusions

This paper described the establishment of a module *Introduction to Data Science* in the full degree bachelor program of Applied Computer Science at HTW Berlin.

It was shown how the module units were designed in order to be able to meet underlying conditions and goals. The module aims both at exposing students to multiple disciplines and at offering students practical experiences in seminaristic lectures, lab exercises supplemented by student team projects.

The presented syllabus includes a broad set of foundations and integrates domain-specific student projects on the basis of a workflow. It was described how-due to a lack of consensus in existing theory-a workflow has been established in order to guide students with their projects. In addition, an approach to a systematic scheme for consistently documenting and reporting data science project work and course grading was presented.

The module has been evaluated by presenting student evaluations and pursuing the question how the established module meets proposed key concepts and skills associated with data acumen. This was done by taking available recommendations by The US National Academies of Sciences, Engineering, and Medicine as well as the EDISON Data Science Framework as a basis.

Although the approaches address data science programs, not single modules, it seems that the module described in this paper is able to serve building foundations associated with data acumen by addressing multiple described key concepts and competencies. These embrace mathematical, computational and statistical foundations, data management and curation, data description and visualization, data modeling and assessment, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving.

The EDISON Data Science Framework proposes two further competencies characterized as highly demanded and specific to data science: 1. Preservation and 2. Scientific or Research Methods [3, pp. 622f.]. The module seems to address various of these competencies.

In addition, students seem to be satisfied with their learning progress based on the student evaluation (cp. Table 3). The module is characterized by a high demand constantly exceeding the planned capacity.

Yet, several limitations can be highlighted. The module does not address:

- Mathematical foundations in terms of graph and network theory. Neither does the whole elective study program at the bachelor's level.
- Data description and visualization with regards to dynamic visualizations and dashboards.
- Communication and teamwork in terms of offering students the ability to understand client's needs within a real industry data project. Topics and domains of projects are chosen by student teams, not by real clients. Therefore, it remains unanswered how well this module or the study program serve a demand from an employer's perspective. With regard to the CRISP-DM framework [10] the module does not address business understanding as the starting point to a data science workflow.
- Data preservation competencies as proposed by the EDISON Data Science Framework [3], while this is the focus of other modules in the study course.

Additional findings can be highlighted as follows:

Student evaluations should be executed more frequently with a higher desired response rate. The questionnaire measures degrees of satisfaction and is only one perspective, neglecting other parameters that should be defined for an evaluation metric.

The module capacity is not able to meet the demand. The planned capacity of the module has been constantly exceeded by the actual high demand (enrollment). This might point toward offering more capacities to the field of data science in the standard curriculum.

Theory and practice lack a consensus pointing towards a general framework to structure the process of solving a problem by means of data science. A workflow model integrating ethical problem solving is missing. Although an approach to a workflow model has been proposed in this paper (cp. section III), this can only be seen as one tailored approach for a teaching situation in view of multiple other approaches. More work has to focus on building a workflow model that might serve both academia and practice while being adaptable to other process models, e.g. (agile) project management and/or software engineering.

Lastly, the approach to assessing the module described in this paper provided several insights and therefore seems useful in order to assess the contents of the module with regards to their adaptability to building data acumen. Nevertheless, a comprehensive metric for assessing and evaluating the work of data scientists in different roles (e.g. practitioner, student, instructor) and domains is needed.

In conclusion, the findings allow a first insight into the challenge of establishing an introduction to data science.

Introduction courses are generally confronted with the need for doing the splits between foundations needed in general, foundations provided by other modules, foundations needed by other modules, demand of domain-specific knowledge and the skill level of both the students and the instructors.

It is hoped that the module presented here contributes to providing students with the prerequisites and the ability to make good judgments, use tools adequately and responsibly, and make good decisions using data. The presented approach to and experience with the introductory course to data science at HTW Berlin reflects one approach that seems to serve data science education. In addition, several spotted limitations justify future work in theory and practice. Further analyses should focus on potential refinements of the module syllabus and the whole study program in order to achieve a constant improvement of the contribution to data science education.

References

[1] J. Bell. *Machine Learning: Hands-on for Developers and Technical Professionals*. John Wiley & Sons, Indianapolis, IN (USA), 2014.

[2] Council of European Union. Regulation (EU) 2016/679 of the european parliament and of the council (general data protection regulation). In *Official Journal of the European Union*. EU, 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN.

[3] Y. Demchenko, A. Belloum, W. Los, T. Wiktorski, A. Manieri, H. Brocks, J. Becker, D. Heutelbeck, M. Hemmje, and S. Brewer. Edison data science framework: A foundation for building data science profession for research and industry. In *IEEE 8th International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 620–626, Luxembourg City (Luxembourg), 2016. IEEE.

[4] Department for Digital, Culture, Media & Sport. Data Ethics Framework. In *Guidance*. United Kingdom Government Digital Service, 2018. https://www.gov.uk/government/publications/data-ethics-workbook/data-ethics-workbook.

[5] D. Dietrich, B. Heller, and B. Yang. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons, Indianapolis, IN (USA), 2015.

[6] N. Döring, J. Bortz, and S. Pöschl. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer, Berlin, Heidelberg, 2016.

[7] J. Janssens. *Data Science at the Command Line: Facing the Future with Time-Tested Tools*. O'Reilly, Sebastopol, CA (USA), 2015.

[8] National Academies of Sciences, Engineering, and Medicine. *Data Science for Undergraduates: Opportunities and Options*. National Academies Press, Washington, DC (USA), 2018.

[9] C. O'Neil and R. Schutt. *Doing Data Science: Straight Talk from the Frontline*. O'Reilly, Sebastopol, CA (USA), 2014.

[10] R. Wirth and J. Hipp. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39, Manchester (UK), 2000. Practical Application Company.