Data Science and Energy: Some Lessons From Europe on Higher Education Course Design and Delivery

Hussain Kazmi^{1,2,3} Ingrid Munné-Collado⁴ Khaoula Tidriri^{5,6}
Lars Nordström⁷ Frank Gielen^{2,8} Johan Driesen¹

¹Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium,

²InnoEnergy, European Institute of Innovation and Technology, Brussels, Belgium,

³Research Foundation – Flanders, Brussels, Belgium,

⁴Centre of Technological Innovation in Static Converters and Drives, Department of Electrical Engineering, Polytechnic University of Catalonia, Barcelona, Catalonia, Spain,

⁵The French National Centre for Scientific Research, Grenoble Alpes University, Grenoble, France,

⁶Grenoble Institute of Technology, Grenoble Images Parole Signal Automatique Lab, University Grenoble Alpes, Grenoble, France,

⁷Division of Electric Power and Energy Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden,

⁸Internet Technology and Data Science Lab, Ghent University, Ghent, Belgium

Published on: Jan 27, 2022

DOI: <https://doi.org/10.1162/99608f92.fd504fc4>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Data science is seen as a key enabler for technologies that help decarbonize global energy use. However, the energy sector continues to struggle to attract and train enough data scientists. The primary reason for this is the lack of emphasis on data science in most graduate programs in energy engineering, and the high barriers of entry for data scientists from other sectors. In this article, we present a snapshot of the data science–related curriculum being taught in graduate energy programs in four different European universities as well as include feedback we received from students and alumni of these programs. While knowledge of data science remains low across the board, students in these programs already recognize data science as an important element of their future professional careers. We also present findings from running three separate iterations of an energy data science course we developed in light of this feedback—one of these iterations was offered only in KU Leuven (Belgium), while the other two were accessible to students at all four universities. In the article, we also discuss challenges and opportunities arising from designing and delivering courses in a cross-university context. This foundational course and others like it are seen as a necessary means to enable students to take more specialized courses in data science, and eventually contribute toward realizing a sustainable energy transition and meeting climate change mitigation objectives.

Keywords: energy, data science, higher education, exploratory data analysis, forecasting, decision making

Media Summary

The energy sector is arguably the single most important contributor to anthropogenic climate change. Mitigation efforts have revolved mostly around decarbonization of energy supply and electrification of demand. However, the vast amounts of data being collected today can enable a more inclusive transition with far less disruption. Yet, despite this tremendous potential, the energy sector finds itself facing a constant shortage of engineers who possess the necessary data science skills. One important reason for this is a graduate-level curriculum that prioritizes technology and policy aspects at the expense of data literacy. Furthermore, there are large gaps in even basic programming knowledge in graduate energy students that hinder them from applying data science in practice. Likewise, as with any other fast-moving discipline, there is a considerable dearth of teaching materials and expertise.

In 2019, EIT InnoEnergy helped set up an international working group comprising participants from four partner universities across Europe (KTH Stockholm, KU Leuven, UPC Barcelona and Grenoble-INP) to help create the blueprint for a new data science program for energy engineers. This partnership was intended to create a targeted data science offering while reducing replication of work and enabling sharing of best practices across universities. An additional benefit of proceeding in this way was to free up instructor time to focus on developing more advanced courses dealing with local research competences.

As part of its activities, the working group evaluated both the curricula and students' current knowledge in four different cross-university graduate programs in energy. The results confirmed our initial hypothesis: students self-rated their knowledge of data science–related skills as low or very low, while considering it important or very important to their future career plans. The curriculum in place in all four programs likewise paid little attention to data science for energy holistically, with only some exceptions.

In light of these findings, the working group has worked on developing course content and delivery methods to cover the most critical elements of data science for energy engineers. Using the specific case of demand response, students have had the opportunity to explore the end-to-end life cycle of data science as it applies to energy projects in practice. The course was offered twice in 2020, once to a cohort of 13 students at KU Leuven, and once in an online setting to 108 students from all participating universities. In 2021, an expanded version of the course was offered as a 3 ECTS (European Credit Transfer and Accumulation System) summer school, open to InnoEnergy Masters students from all over Europe. The course saw encouraging levels of engagement from the students, and students found some elements, such as the extensive use of Jupyter Notebooks and the end-to-end case study, particularly helpful. At the same time, we also unearthed a number of practical challenges in the design and delivery of this course which form the basis for our ongoing collaboration.

1. Introduction

The global energy sector is undergoing a profound transformation as part of accelerating mitigation strategies to meet climate objectives. On the energy production side, renewable energy sources such as wind and solar are gaining in popularity. However, unlike conventional fossil-fuel based power plants, the output of such renewable sources cannot be arbitrarily modulated to match consumer needs. With energy storage still an expensive proposition, increasing proliferation of renewables necessitates peaking power plants in many places, often in the form of polluting gas or diesel generators. At the same time, on the consumer side, energy demand is projected to grow rapidly in non-OECD (Organisation for Economic Co-operation and Development) countries due to greater electrification, proliferation of cooling equipment, and climate change (van Ruijven, 2019). It is also set to grow in OECD countries (or at least offset demand reductions due to efficiency measures) because of the rise of electric heating and mobility (Blonsky, 2019). These trends are shown in Figure 1, which is adapted from a recent report on grid adequacy by Elia, the Belgian transmission system operator (Elia, 2021). Meeting this increasing electricity demand using fossil fuel–based power plants will further exacerbate climate change problems.

Figure 1. The expected evolution of the global energy system in the coming decades.
Adopted from Elia (2021).

These developments have been accompanied by a greater focus on digitalization, including extensive use of sensors, for example, smart meters, besides other sources of data and ICT (information and communication technology) tools. In fact, today, extensive data about energy demand and production is collected and processed at different scales of the grid in real or near-real time. Alongside supply-side decarbonization and electrification of demand, using gathered data to model and optimize energy flows are seen as a key enabling technology for a more efficient system. In addition to helping optimize grid operation, such ubiquitous sensing also brings other benefits: enabling, for instance, improved asset management, and fault detection and prediction. The data collected in this way can also inform policy decisions on long-term operational planning, and help engage citizens to become more conscious consumers of energy (Kazmi, 2021).

The energy transition can therefore be summarized along six dimensions: decarbonization (of energy supply), digitalization (through gathering and storing relevant data), decentralization (of generation and storage, e.g., through battery systems), disruption (of business-as-usual supply centric energy system), democratization (through consumer and prosumer participation), and demand increase (due to socioeconomic changes, as well as electrification of heating, cooling, and transportation, etc.). The importance of this energy transition can be hardly overstated in the face of overwhelming consensus on anthropogenic climate change in the scientific community. Each of these six dimensions consequently has a role to play on its own, but they must come together to create a single ecosystem. Data science, or evidence-based decision-making in general, as we will show in a subsequent section, will provide the glue required to build this ecosystem.

Yet, data science has remained almost wholly absent in energy-related curricula across Europe, and indeed most other places. This is in stark contrast to workplace requirements, where data-related skills have clearly become a necessity to make the energy transition happen. Consequently, graduates from energy engineering programs joining the workforce today have large gaps in their data literacy, and as such are often required to either learn on the job, or make use of outdated programming languages, tools, and skills. Companies continue to suffer from this lack of skills, with many lacking even basic automation and visibility due to a continuing reliance on spreadsheet software such as Excel.

In the recent past, there have been a number of isolated efforts to address these shortcomings. Some of these have been specifically focused on energy education (Hong, 2018; Pathak, 2016), while others have been more generally aimed at STEM (Science, Technology, Engineering and Mathematics) education (Bybee, 2010; Lue, 2019). Some universities, such as University College London, Exeter University, and Offenburg University, have developed entire one- or two-year MSc programs on energy data science. These typically cover a broad spectrum of relevant topics, but are usually seen in the narrow context of a standalone program rather than integrating with the broader energy engineering education. Other universities, such as National University of Singapore, have likewise developed a number of modules on applying data science in the built environment, however, the focus in many such courses is not on the holistic energy system, which also includes grid-side considerations. A number of similar initiatives have emerged in recent years in North America and other parts of the world. However, since there is no inter-university coordination in the development of almost any such program, they remain fragmented, and replication of instructor work is frequent.

A number of practical initiatives, such as the Data Science Connector courses from UC Berkeley, have attempted to make data science accessible to different fields by fostering communal collaboration in both teaching and research (Adhikari, 2021). These courses also shine a light on how computational and inferential thinking must be interleaved for an effective education in data science: either component on its own is arguably not sufficient for students to fully leverage the advances in the field. Such collaborative programs that allow sharing of best practices and learnings, ranging from course design to delivery, do not exist to the best of our knowledge in the narrower domain of energy data science.

More generally, even though most educational programs on data science are quite recent, there is a large body of literature on innovations and challenges in STEM education that can be drawn upon to develop best practices. Over the years there has been significant debate on what this entails in practice, but there is sufficient consensus in the field to (1) emphasize integrated approaches to teaching and learning, where discipline specific content is treated as ‘one dynamic, fluid study,’ rather than being divided and treated separately, and (2) further investigate integrative STEM education in a way that helps teachers understand the effect of teaching methods on student engagement, and so on (Brown, 2012). Moreover, based on a number of studies analyzed in a recent survey, it is evident that the integration in STEM education is most commonly associated

with the engineering discipline and real-world contexts (Martín-Páez, 2019). This is a theme we will revisit repeatedly in our own course design as well.

To reduce the gap in energy data science education while keeping in mind broader STEM education trends, EIT InnoEnergy helped set up a working group constituted of members from universities in Belgium (KU Leuven), Sweden (KTH), and Spain (UPC) in 2019. This was later expanded with a university from France (Grenoble INP). EIT InnoEnergy is one of the knowledge and innovations communities (KIC) of EIT, the European Institute for Innovation & Technology, with a focus on the sustainable energy transition. Over the last decade, EIT Innoenergy has been setting the stage for the EU Green Deal with investments in innovative projects, startups and talent. Its educational arm has seen more than 1,200 MS students and 260 PhD students graduate from eight different programs across Europe over the last decade. A hallmark feature of InnoEnergy MS programs is mandatory mobility, whereby students are required to study in at least two European universities over the two-year program. This requirement not only allows students to utilize resources available in different universities, it also allows participating universities to harmonize their course offerings while further developing local competences at the same time. This applies even more to courses and programs that are now under active development—such as ones in data science.

The idea behind setting up the data science working group was therefore multifold. A key objective was to improve communication across universities and countries partnered in different EIT InnoEnergy programs to avoid duplication of work, create a consistent curriculum, and propagate best practices. This meant that the core foundations of data science should be distilled into a replicable course template, whereas local expertise at each university would then build on top of this foundational course. The foundational course developed in this way was meant to provide students of varying backgrounds and skills with a practical, hands-on introduction to data science. Furthermore, the working group was also tasked with creating a future roadmap for graduate programs in energy engineering. This is a challenging activity in and of itself, since each university has multiple programs that include different students where similar content is covered. Another interesting consideration in this curriculum development was how best to bridge the gap beyond industrial requirements and academic state of the art in a rapidly evolving discipline, something that is of grave concern in STEM education in general as well.

In this article, we document key findings of the working group so far, and share some key lessons learned while offering a course on energy data science over three iterations in very different settings. We also highlight some learnings and evolving practices from sharing knowledge across different universities and programs. As it is a heavily overloaded term, we begin by defining the scope of energy data science.

2. Data Science in the Energy Sector

We take data science to refer to the broad set of tools that enable practitioners and researchers to make use of data and domain knowledge to arrive at decisions. Energy data science narrows down this scope of application

to energy systems. Today, it finds applications in all aspects of modern energy systems, holding true for both upstream and downstream operations on the supply and demand side. In this section, we discuss some common use cases and identify which ones have been explored in the developed curriculum, and which are candidates for subsequent addition.

The oil and gas sector has arguably been one of the earliest adopters (and innovators) of data science principles to detect and identify hydrocarbon reservoirs (Febowitz, 2013). This frequently makes use of various data sources ranging from accurate but sparse geological (well log) data to abundant but difficult to interpret geophysical data. The challenge for the data scientist in this domain is to integrate information from multiple sources (geophysical, geological, etc.) with domain knowledge to optimize hydrocarbon production while minimizing risk. Challenges surrounding visualization and how to deal with big data abound.

As a program rooted in sustainability, the use cases we explore in the context of our working group are a bit removed from fossil fuel operations. More concretely, accurately forecasting energy generation using renewable energy sources and addressing emerging challenges to grid operation due to their accelerating proliferation are interesting data science problems (Sweeney, 2020). Perhaps not so surprisingly, many techniques developed in the context of oil and gas exploration find modern analogs in using data to determine the best location for wind and solar photovoltaic (PV) energy. Techniques such as kriging and large-scale reservoir simulations can allow planners to optimize the location of wind turbines and solar plants. Likewise, once in operation, data science techniques can be used to ensure optimal power production from these resources. Any deviations from expected power output can be monitored using, for instance, anomaly detection algorithms. Faults can be detected and, in the best case, predicted to minimize downtimes and maximize capacity factors.

In a similar vein, data science tools are widely used by grid operators to make timely decisions on how to optimally operate the electric grid when faced with increasing amounts of these variable renewable sources such as wind and solar. This allows grid operators to effectively reduce the carbon footprint of the energy sector by only relying on fossil fuels when necessary. The carbon footprint data for a large number of locations is available from system operators or through third parties such as electricitymap.org, which can be useful for short- and long-term planning. ENTSO-e, an umbrella association for European Transmission System Operators, provides both historic and future forecasts for different energy sources in the grid as well as energy demand for most European countries through its transparency platform. These datasets can be used to quickly build a better understanding of the current energy mix and its sustainability in a large number of countries across the world.

Finally, data science is also being increasingly employed in demand-side management programs, either through automation or through engaging end users and providing them with feedback to minimize their carbon footprint (Kazmi, 2020). While engaging users creates a bottom-up movement of citizen empowerment, which ties in well with Europe's focus on citizen energy communities (Sokołowski, 2020), automation is likely to

provide the largest gains in using demand flexibility to help stabilize the grid. Finally, it is important to mention that the synergistic relationship between data science and energy has evolved considerably in recent years. From energy engineers using data science to reduce the energy consumption of consumers and data centers (Uddin, 2011), the focus has slowly evolved to encompass green-ML (machine learning) methods, which ensure that ML models are deployed in an eco-conscious manner, keeping in mind their potentially enormous footprint (García Martín, 2017; Strubell, 2019).

The easy availability of energy data from numerous online sources means that it can be correlated with real-world events to show students their impact on energy supply and demand. Some obvious examples of this includes the impact of an eclipse on solar production and the EURO football competition on electricity loads. In light of these applications, we shortlisted three skills as the building blocks for evidence-based, data-driven decision-making that underpin energy data science: (1) programming skills, (2) machine learning knowledge, and (3) optimal decision-making expertise. We assume that machine learning knowledge also includes data wrangling, visualization skills, and probability and statistics. Optimal decision-making, in the context of energy engineering, refers more to optimal control and design challenges, but it can also include using optimization algorithms for training ML models and optimizing their hyperparameters.

3. Data Science Education Prior to Commencement Of Activities (2019)

In this section, we present a brief overview of the status of data science education and students' knowledge in the energy programs under consideration prior to the working group initiative.

3.1. State of Curricula

InnoEnergy administers eight different MSc programs. Each of these MS programs has a different focus, and some are more amenable to data science techniques than others, as we will show in this section. All of these programs are offered at a number of different European universities, mostly ranging between four and seven per program. In many cases, the same university is involved in multiple MSc programs, and students from one program can occasionally take a course in a different MSc program. For the course of this project, we shortlisted four such MSc programs (SELECT, RENE, SENSE, and Smart Cities) to analyze the state of data science curricula in different universities, which include KU Leuven (Belgium), KTH (Sweden), Ecole Polytechnique Paris (France), TU/e (The Netherlands), Politecnico di Torino (Italy), Universitat Politècnica de Catalunya UPC-BarcelonaTech (Spain), INP Grenoble (France), and IST Lisbon (Portugal). Besides the host universities, these programs also differ considerably in their student intake. For instance, SELECT is more oriented toward mechanical engineering, as opposed to SENSE, which has more of a focus on electrical engineering topics. The fact that each of these programs accepts students from all over the world from different undergraduate fields of study also contributes to their diversity. Of the universities offering these four

programs, KU Leuven, KTH, UPC, and INP Grenoble were represented in the working group. The results, as they stood in the 2019–2020 academic year, are presented here:

1. At KU Leuven, students in the selected programs had the possibility to take a course in smart distribution systems, which covered a wide spectrum of topics ranging from sensing to automation and machine learning to control. More concretely, in one of the lab sessions, students were required to train a neural network to predict electricity spot prices. However, besides this course and the very general introduction it provided, there was no specific course that addressed data science for energy. Students could optionally take courses such as industrial control and automation, optimization, building simulations, and experimental design, but most of these were in general settings, and either not targeted at the energy domain or not adequately focused on data science.
2. At KTH, students could study a course on communication and control in electric power systems and an additional course on computer applications in power systems. However, these courses provided students only with an introduction to machine learning and how it could be applied to power systems providing exercises in Java. Furthermore, in two of the four programs we considered, these were optional courses and many students ended up not taking the course. A later analysis showed that these students were in fact the ones that were most in need of such a course.
3. At UPC, students in one out of the four programs could take a course titled Control and Automation for the Efficient Use of Energy, which introduced them to programming Arduino and Raspberry Pi boards, and using these to control flexible household appliances such as the washing machine, the EV charger, water boilers, or space heaters. However, the focus of this course was also not on data science skills per se, but rather on creating an end-to-end application with simple controllers. As with KTH, this course too was optional for three out of four programs, therefore most students in these programs ended up not taking it.
4. At INP Grenoble-ENSE3, since 2018, students can choose to follow a ‘digital journey’ over three years, with one data science–related course for each year. This digital journey aims to explore all stages of data processing ranging from acquisition and storage to modeling, analyzing, and securing data. This is achieved through lectures, practical work, and challenging projects. Examples of courses include Data Science Initiation and Machine Learning and Optimization. However, since the digital journey is open to all engineering students, the courses do not focus specifically on data science for energy systems but cover many applications (intelligent systems, power systems, mechanical, hydraulics, and civil systems, etc.). Furthermore, this digital journey is still optional, and therefore suffers the same fate as in other universities.

From this summary, it became quite obvious that data science–focused courses for energy engineers were mostly absent in KU Leuven, INP, and UPC. Based on anecdotal evidence, the situation in most other European universities is largely similar. Even when general purpose data science courses exist, energy students tend to steer clear of them. One reason for this is that students in energy programs lack the programming skills that many such general-purpose courses require. Furthermore, many students are simply unaware of the options available to them (which are often administered by a different department), or do not understand how

these courses can form part of a data science education. This also mostly explains why students did not opt for existing data science–related courses in most programs. Unsurprisingly, in most energy programs, optimization and optimal decision-making is given more importance, often at the expense of the machine learning and data literacy aspect of a full-fledged data science course. This means that countless engineering graduates are entering the workforce without ever having used a programming language besides MATLAB, let alone possessing the skills necessary to thrive in a modern data science team.

The information presented in this section was gathered through regular meetings of the working group. However, the unstructured form of the questions and discussions meant that, while a lot of knowledge was shared, it was difficult to scale the process (to other programs and universities, as well as in time). In a subsequent section, we discuss the current means to share this information, which trades off some knowledge for structure.

3.2. State of Students' Knowledge

Concurrently with analyzing the current curriculum, we also approached students studying in these programs directly to better understand their current data science–related skills in the three categories we discussed earlier: (1) programming skills, (2) machine learning knowledge, and (3) optimal decision-making know-how. Based on survey responses from over 150 current students and alumni in the four InnoEnergy programs, we drew the following conclusions:

1. An overwhelming majority of respondents (75%) reported having poor or very poor knowledge of data science as a field, while ranking it important or extremely important to their future career. This trend has held between 2019 and 2021.
2. A slightly lower fraction, but still more than 50% of the respondents, self-rated their programming knowledge as average or below average. Between 2019 and 2021, we have seen a gradual, albeit small, improvement in self-reported scores in programming, which may be reason for cautious optimism.
3. Second-year students in all programs consistently self-reported better knowledge on the three metrics (programming, machine learning, and optimization); this effect was the largest in optimization, followed by programming and machine learning.
4. Students in energy, mechanical, chemical, and civil engineering consistently self-rated worse than electrical engineers on all three metrics. Owing to the nature of student intakes in the different programs, this also meant that prior student expertise was unevenly distributed.

In light of the current curriculum, none of these conclusions came as a huge surprise. The students' perceived importance of data science and their motivation to study it mirrors industrial professionals' opinions. For instance, in an independent survey of over 100 energy utilities, professionals and managers considered energy forecasting, data analytics, and automated energy trading among the most important skills for the future (Hong,

2018). Student opinions of the most important topics reflected these as well. This information made a considerable contribution to the design of the foundational course, as we discuss next.

4. Course Design, Iterations and Intended Learning Outcomes (ILOs)

In light of the above discussion, it was obvious that there was a need for a specific energy data science–related curriculum that was open to students from all MS programs, irrespective of the university. The course, since its inception, has therefore had a number of clearly defined intended learning outcomes. These include that students should be able to (1) access energy data sets from a variety of different sources, and perform exploratory analysis on these data sets; (2) create models explaining the patterns they observe in the data, and predict them in the future; (3) use these models in optimal decision-making frameworks to improve control or design of energy resources; and (4) appreciate the practical challenges of deploying these solutions at scale. In addition, it is important for the students to realize the importance of always benchmarking ones’ results against simpler alternatives, as well as the idea of significance testing when comparing the performance of different algorithms. These ILOs were chosen based on the results reported in Hong (2018) as well as interactions with European academics and energy companies. However, the primary focus during the course is on an understanding of the tools rather than the context, which can be exchanged for a different use case without loss of generalization.

Consequently, the course was trialed initially at KU Leuven in spring 2020 for MSc energy students, before being opened to all InnoEnergy students in summers of 2020 and 2021. Figure 2 provides an overview of the evolution of the course, along with the students that enrolled in the course and those that attended it (this was only different for the case of the latest iteration in 2021 where we applied a basic shortlisting procedure in the form of a motivation survey for attending the course). The constitution of the students attending the course has also evolved considerably. While the second iteration was open to InnoEnergy MSc students as well as alumni and PhD students from a number of different partner universities, this was limited to only currently enrolled MSc InnoEnergy students (from any partner university) and KU Leuven PhD students.

Figure 2. An overview of the three course iterations.

The coverage of material has also greatly expanded in this period, with the assigned credit roughly doubling (from 1.5–2 ECTS in the first iteration) to 3 ECTS in the latest one. However, the approach has remained largely consistent: the course was structured to follow a flipped classroom approach with five lectures on the theory of data science. These lectures were followed by a practical lab session in which students worked with Jupyter Notebooks provided to them. The different course modules made use of a single data set throughout, which consisted of electricity demand in a neighborhood of 200 different houses over an entire year, sampled at 15-minute intervals (Labeeuw, 2013). For the final session, students were also provided real-time electricity prices from Belgium. The individual sessions are organized as:

1. The first session provides students with an introduction to energy data science as well as useful pointers to brush up on their programming skills using existing notebooks taken from the University of Cambridge's excellent introduction to Python programming.
2. The second session on exploratory data analysis allows students to explore the provided data set in greater detail using summary statistics, filtering, and resampling. They were also able to extract general trends in electricity demand as a function of time of day and day of year here. Additionally, visualization tools and techniques to deal with missing data were also introduced at this point.
3. In the following session on modeling and forecasting, students explore time-series analysis concepts such as correlations (auto-, partial auto-, and cross-correlation), as well as time-series decompositions into trend, seasonality, and noise components. Using demand data from the previous session, students are then able to create forecasts using three different families of algorithms for an entire year (baseline persistence, multilinear regression, time series methods). The data also allows students to better understand the relationship between the accuracy of forecasts at different demand aggregation levels, ranging from the individual household level to the neighborhood level.
4. In the optimization and optimal decision-making session, students use the demand data from previous sessions and time-varying electricity prices from the Belgian spot market to optimize the charging and discharging profile of an electric battery. Students are required to use the best forecast they developed in the

previous session with the optimal control to better understand the monetary costs of misforecasts and see how they could further improve their decision-making using improved forecasts.

5. The final lab session was followed by a lecture that covered advanced topics such as data quality in the real world, privacy-aware learning, comparing different algorithms in a statistically meaningful way, and how to use transfer learning to mitigate lack of data in practice.

Each of these topics can easily form the basis for a semester-length course, therefore we had to make a sampling of the most important aspects of each. In the first iteration, students were evaluated on their participation in a group project where they were required to come up with the optimal design of a battery-inverter system. This made use of the concepts introduced in the course (exploratory analysis, forecasting, and optimization). However, a number of changes introduced in 2020 meant that the second iteration brought an increased focus on the computational aspects of data science. Some of the additions in the course content included an introduction to techniques and data formats that can be useful in working with large data sets. Furthermore, an increased emphasis was placed on Pandas (a Python library), as a means of showing students how to work with time indexes to resample, filter, and aggregate time-series data. A number of additions were also made to the visualization techniques included in the course. Students successfully finishing the course would have worked with a considerable number of data science-related libraries. The course project was replaced with a Kaggle in-class competition, where students had to create a single, week-ahead forecast for electricity demand in an aggregation of buildings.

Likewise, the third iteration of the course further increased the emphasis on the inferential aspect of data science. More specifically, besides introducing statistical significance in comparing performance of machine learning algorithms, we adopted an optimization-centric view to the lectures on machine learning and optimal decision-making. In the first step, in both topics, students are shown how well random guessing or sampling can work at estimating the parameters of linear models as well as controllers for battery storage systems. While the first is a popular trope in introductory ML lectures, the latter is unfortunately something we have also witnessed in the industry. These concepts were followed by a description of gradient-based and gradient-free techniques, which can solve the problem in a better way than random guessing. Finally, techniques that are guaranteed to converge to the optimal values are introduced (e.g., the normal equation to estimate parameters of linear regression, linear programming to optimally solve the battery control problem). The intent behind such a step-by-step approach is to make the students aware of the close connections between machine learning and optimization. It also allows us to introduce concepts such as computational complexity, and how nonconvex optimization problems cannot be solved, without modifications, with exact but convex solvers. This changelog from the first to the third iteration is highlighted in Table 1.

Table 1. An overview of the course changelog from the first to third iteration.

Iteration 2	Iteration 3

Introduced a number of new libraries: Plotly for visualization, Modin for speeding up, Nevergrad for optimization.	Introduced Pulp for optimal decision-making as a way to contrast it against derivative-free and simpler optimization schemes.
Expanded theoretical content (correlational analysis, many different classifications of forecasting and optimal decision-making algorithms as they relate to the energy domain, advanced concepts such as transfer and privacy-aware learning, etc.).	Expanded theoretical concepts (time-series decompositions, state of the art forecasting techniques from M competitions, improved benchmarks for optimal decision-making, etc.).
Replaced design competition with Kaggle forecasting competition.	<p>More emphasis on inferential aspect of data science, including testing hypothesis and comparing forecasters statistically.</p> <p>Practitioner talks to illustrate how data science is actually used in practice.</p> <p>Further developed the Kaggle competition and introduced an open-ended exploratory data analysis assignment.</p>

In addition to these conceptual adaptations, the latest iteration also includes a section on advanced modeling techniques with a specific focus on addressing challenges typically faced in the energy domain. These include a discussion of local vs. global modeling techniques, which provide an interesting view of scalability in the real world where we are typically interested in forecasting hundreds or thousands of time series in near-real-time, showing students a view that bridges inferential and computational thinking. These techniques are then used to introduce more advanced concepts such as transfer learning and auto-ML. Furthermore, with the latest iteration, we have decided to host at least two practitioner talks. This includes a top-down presentation from an aggregator of energy demand, which provides the big picture view. The second talk, from a practitioner at a service provider company, builds from the ground up and takes students through the model deployment process, that is, how to deploy machine learning and optimization models in practice. In addition to making the Kaggle competition more realistic in the latest iteration, a new assignment on exploratory data analysis was introduced where students are asked to tell a story using an energy data set of their own choice. The Jupyter Notebooks used in the course are available at <https://github.com/hussainkazmi/energyDS> under an AGPL license.

5. Course Delivery

While the first version of the course largely took place in an on-campus setting, the subsequent iterations have been fully online. The 2020 summer edition was organized as a series of six lectures, spaced one week apart. Each lecture would last roughly one to one and a half hours, and cover theoretical elements. In the ensuing week, the students were required to go through the provided Jupyter Notebooks and complete exercises therein, which would subsequently be discussed. This long format proved to be rather hectic, both for the students and

instructors, leading to a high dropout ratio. Consequently, in 2021, we switched to a more condensed, two-week format. In this version, the first week was wholly dedicated to the theoretical lectures and work on the notebooks. Based on student feedback from previous iterations, the theory lectures were expanded and lasted over two hours long (in the mornings). In the afternoons, there were open office hours for students to discuss any questions regarding theory or practice (or the provided Jupyter Notebooks). The second week was dedicated entirely to course readings, interactive question and answer sessions, and the Kaggle competition.

The Jupyter Notebooks, which form an important part of the course, have seen different hosts over the three iterations. For the first two iterations of the course, we used a custom-built cloud platform based on JupyterHub that allowed students to run their Python scripts on the cloud. With all the packages preinstalled, this removed a frequent point of failure for new Python users, smoothing the learning curve. Students could now focus on developing their algorithms. The cloud platform also granted students greater computational resources than available on a typical laptop, and therefore allowed students to explore implemented algorithms in greater detail (or implement their own from scratch). From InnoEnergy's perspective, this also gave the instructors greater control over content. In the third, most recent iteration of the course, we have migrated to Deepnote, a cloud platform tool for collaborating on Jupyter Notebooks. This was easier (and cheaper) to maintain than a custom implementation, and provided a lot of the functionality we needed to successfully run the course out-of-the-box.

To get students up to speed with Python programming, we have made use of two existing resources. To start with, students are provided an introductory “Lecture 0” with Jupyter notebooks taken from University of Cambridge’s excellent self-study introduction to computing with Python (<https://github.com/CambridgeEngineering/PartIA-Computing-Michaelmas>). Additionally, till last year, we also provided interested students with licenses for DataCamp, an online resource for learning programming and foundational data science skills. This was only used by a few students so it was largely discontinued in 2021. Students could also ask Python programming–related questions on the course forum or during open office hours. The level of Python knowledge required for the course was also intentionally kept rather low (i.e., object-oriented programming, creating packages, and other, more advanced concepts are not covered).

6. Outcomes

In contrast with the previous section, which focused on what we taught, this section highlights some of the key things we learned while developing and delivering the course.

6.1. Evaluation of ILOs

To evaluate how well the course does in meeting the ILOs, we introduce an assignment on story-telling (i.e., exploratory data analysis on a dataset of students’ choice) as well as a Kaggle competition on forecasting energy demand in a group of buildings given both historical data and some exogenous variables. The assignment on exploratory data analysis is open-ended and the students are expected to choose their own

energy-related data set for analysis. They then receive feedback on how the analysis can be further improved or how it would be carried out in practical settings. Most students were able to choose an interesting data set, and run exploratory analysis on it, leading to a coherent report on what the data set could be used for in practice. These data sets ranged from energy demand to generation, but some were more obscure as well, dealing with the performance of electric vehicle batteries, and so on.

The Kaggle competition, on the other hand, is a group activity where participating students implement a large number of different models (ranging from linear models and Facebook Prophet to gradient boosted trees and neural networks) as part of the challenge. The basic requirement for students to pass the course was to receive a pass grade on the exploratory data analysis assignment and to beat a simple persistence model on the forecasting challenge. All groups managed to achieve this latter target (albeit with varying margins). In the final retrospective, this issue of different forecast accuracies was emphasized, along with the concept of statistical significance and hypothesis testing when dealing with many models for the same phenomenon. While a poor model may occasionally perform better than a better model, it will fail to do so consistently.

6.2. Student Feedback

The reception of the course has ranged from positive to very positive, with the most recent iteration also receiving the most positive feedback yet. In this section, we provide quantitative and qualitative assessments of the course with a focus on the latest iteration. Where applicable, these results are also contrasted against previous runs.

Reassuringly enough, around three-quarters of students providing feedback said they were very likely to recommend the course to their class fellows or former selves ($n = 34$ in 2021, $n = 16$ in 2020). This was also reflected in student experience with the course, which has earned a median rating of being very useful consistently—4 on a 5-point Likert scale in the 2020 iteration ($n = 16$), and 8 on a numerical scale ranging from 1 to 10 used for the 2021 iteration ($n = 34$). A vast majority of students (80 – 90%) found the course readings very useful and felt that the lectures provided them with enough theoretical background to data science. More than half the students (both in 2020 and 2021) felt that the pace and level of the course was just about right. On the course delivery side, most students rated their experience of the interactive notebooks very positively (median rating of 9/10, $n = 34$). Both the Kaggle competition and the Deepnote platform were likewise rated positively (median rating of 8/10, $n = 32$).

Students also found the practitioner talks useful (both talks received a median score of 8 on a scale of 1 to 10, $n = 32$). These also generated valuable discussions during the sessions. One thing that we felt provided positive reinforcement was that both speakers were alumni from different InnoEnergy programs, and students could relate to their experiences in the field. This led to a fruitful discussion on transitioning from being an energy engineer to an energy data scientist in the industry, and what such a shift entails.

Despite the largely positive reviews, considerable differences in student experience remained. Almost 45% of the students in 2021 found the course either too basic or too advanced (split almost evenly between students who found the course a bit too advanced and those that found it a bit too basic). This number was about the same as the feedback we received in 2020. Based on the huge diversity in the cohort, this is expected, but something we hope to work toward reducing, possibly through optional preparatory lessons at the beginning of the course. Likewise, a considerable number of students felt there was either too much theory in the course (2021) or too little theory (2020). This delicate balance between theory and practice is something we will continue to strive for.

In subsequent discussions with students, it has become obvious that students have been able to incorporate a number of things taught during the course. For instance, two MSc students engaged in thesis work mentioned that the course would have been useful for their thesis. Based on student feedback, it is also clear that there were some misplaced expectations on the students' side. For instance, some students had very specific use cases in mind, which were not the target of the course. Some of these students got enough relevant material during the course for it to still be useful. For instance, two different students signed up for the course to learn more about modeling and forecasting of electricity prices and solar electricity generation. However, while these were not explicitly discussed during the course, there were relevant readings on the topic; furthermore, the principles discussed in the class and the competition could easily generalize to cover this case. A case where this did not work out was for a student who attended the course hoping to learn about databases, and how to store and retrieve data from them. This topic was not covered, even tangentially, in the course.

6.3. Engagement

Influenced by multiple factors, we have had varying levels of engagement in the three iterations of the course. In the first on-campus iteration with 13 KU Leuven students, all the students passed the course. Due to the cohort being small, each student also received personal attention and feedback on their project and notebooks. The second iteration, open to all InnoEnergy alumni and students, began with a cohort of 108 participants. As there was no academic certificate attached to the course, students attended it on a best-effort basis and there were many dropouts. From 108 registrants, we had around 70 students in the first session, which went down almost linearly to around 20 by the final session. Around 11 students participated in the Kaggle competition and between 20 and 40 completed the tasks in the Jupyter Notebooks. Based on discussions with students and alumni, the high levels of dropouts resulted from several factors. Foremost among these was that the course took up six weeks of the summer, was offered for free, and required substantial amounts of work. Without the carrot of academic credits, only highly self-motivated students who had uninterrupted availability for six weeks managed to finish the course.

This motivated us to make three organizational changes for the 2021 iteration. First, students successfully finishing the course were to receive a 3-ECTS certificate from KU Leuven in 2021. Second, the course was shortened from six weeks to two, to allow students to still enjoy their summer break and get on with other

tasks. Finally, students were required to motivate their participation in the course before signing up. This final requirement, while nothing especially demanding, trimmed the number of attendees to 61 students from the 91 that initially signed up. Of these 61, between 38 and 50 participants showed up during the five sessions in the first week. Of these, 33 students successfully completed all requirements for the course (including participation in the sessions, assignments, and Kaggle competition). This marked a considerably higher retention rate, despite still being an intensive summer course. For future iterations, we plan to make the selection process even more rigorous to further reduce the number of students dropping out between signing up and starting the course.

7. Recommendations and Future Outlook

In light of our experiences, we have a number of broader recommendations for other energy data science educators in particular and data science educators in general. We discuss these in this section.

7.1. Focus on Real Contexts, Hands-On Practice, and Instructor-Led Coding Sessions

Students found the interactive notebooks extremely helpful in understanding both the concepts and how to code, even (especially) when they did not have a Python programming background. However, due to the large diversity in student backgrounds, we recommend a crash course in Python (or a different programming language of choice) preceding the main course to allow everyone to start the course at roughly the same point. There will always be individual differences in students' aptitude to programming and problem solving, but this should not be conflated with an energy data science course, especially if there is only one such course being offered to students. We also found the online cloud platform to considerably help students unfamiliar with programming environments. It was reassuring to note that more skilled students were able to take the provided notebooks and set up their own environments (e.g., on Spyder or Pycharm, etc.). Another helpful feature of the course was the use of an end-to-end real-world example that allowed students to fully grasp the challenges and opportunities inherent in modern energy systems. This combination also reflects the finding from Martín-Páez (2019), which shows that integrative STEM education is very often a combination of engineering discipline with a real context.

A recurring theme in student feedback has been to use instructor-led code-along sessions as part of the course. While we have not trialed this yet, we intend to try it during the next iteration of the course.

7.2. Collaborate, Collaborate, Collaborate

Data science is a fast-paced discipline, and energy data science is no exception. Developments in the past few years have radically altered the face of the field, with advancements in transfer and domain-informed learning techniques providing greater impetus to the widespread adoption of data science techniques. The process of creating a curriculum and long-term vision for energy data science from scratch has benefited greatly from a

collaboration across multiple MSc programs and universities. However, it remains unclear how this collaboration should be structured, and the role individual instructors should play in it. On the one hand, in many ways, working groups such as this one also need to go further than discussing curriculum toward jointly developing and maintaining course material, and ultimately test-beds and platforms. While international organizations such as InnoEnergy certainly have a part to play in these developments, the onus is also on academics to take them further.

As a result, to keep track of the ongoing developments in data science education at partnering universities, we have recently adopted a visual schema that requests course designers and instructors to mention the salient data science topics they intend to teach, along with the specific energy use cases and data sets they intend to tackle. This information is supplemented by target student audience and unstructured intended-learning outcomes. As we mentioned earlier on, such an unstructured format provides us with a lot of information that is, on occasion, difficult to reconcile across universities and instructors. Bringing some structure comes at the cost of simplifying things, but allows quick aggregation of data to see (1) which data science techniques have not been covered and (2) whether there are important energy use cases that still need to be discussed. As a practical example, this schema has quickly helped us identify that a number of partners have already addressed topics such as time-series forecasting and classification for a number of different use cases (demand-side management, fault detection in wind turbines, and electrical power systems operation). This can lead to consolidation, but also minimize duplication of future work. At the same time, the schema helped identify that none of the partnering universities is dealing with some important data science topics such as dashboarding, model deployment, and experiment tracking. This visibility makes such issues easy to address in the medium or long term.

7.3. Transparency and Open-Source Materials Can Greatly Accelerate Collaboration and Adoption

Despite these partnerships, numerous challenges remain. Foremost among these is to harmonize data science education across the different (participating) universities. This is a tall order, particularly since it is something of a moving target; that is, in the years since we set up the working group, at least two to three new data science-related courses have been created in just the universities participating in the working group. One way to address this is through greater transparency and making use of tools such as GitHub, which can be used to share and collaborate on notebooks and course material. As part of the working group activities, we plan to create a shared open-source repository with the code (and course materials) from all partnering universities.

Adoption of courses, once they have been designed and deployed, can also be tricky because in many cases, a university or group may want to develop its own course. This of course remains the prerogative of academics, but organizations such as InnoEnergy can help focus these efforts in directions where they are needed the most (i.e., in specialized and targeted course offerings, rather than in the creation of redundant introductory MOOCs on data science). One challenge that has come up in these discussions so far is licensing and intellectual

property rights of courses created in such an international setting. Even when the courses can be licensed freely (or with attribution), there is still the unaddressed issue of the teacher's role in teaching a course that has already been created. These are open questions that will need to be addressed by the broader data science community in the future.

7.4. Cross-University Partnerships Can Enable Education Testing and Research at Scale

Our intent with course design is for all developed courses to be in the form of largely self-contained modules that can be taught in different programs with only minor modifications. To trial this, for the next iteration, we intend to run a cross-European summer course on energy data science in 2022. While the modalities for its organization are still under consideration, there is consensus that the summer course will be a flagship event open to all InnoEnergy MS students (and perhaps energy students at the partner universities). During the course, the students will work on concrete challenges from the industry in addition to receiving an education in data science. This course will serve as a precursor to develop this existing offering into a more holistic program that covers even more areas of energy data science, including the soft skills necessary to discuss analysis results with technical and nontechnical stakeholders.

Running the same, or a very similar, course in multiple locations at the same time also opens the door to education research on a large sample of energy engineering students. This would not be feasible otherwise if the course was run at a single university: the proposed course addresses the challenge of obtaining large enough samples of (largely) independent but identically distributed cohorts.

8. Conclusion

Over the next decade, we are expecting millions more distributed energy resources such as photovoltaic arrays, electric vehicles, and smart thermostats to be connected to the grid globally. Along with renewable generation, greater control of these devices will be a key enabler of decarbonization to meet climate objectives. But these devices add another layer of complexity to the grid, and, as a massive data network, also create countless points of potential weakness. The energy sector needs data scientists to not just operate these devices in an optimal manner but also help armor the broader network against hack and attack.

Energy data scientists continue to be in short supply and, despite its substantial and pressing needs, the sector continues to find itself in competition with other industries to attract and retain a skilled workforce. It is imperative that new routes into the role are opened up to help fill the void. We hope that the developments initiated here will not just train energy engineers to be data scientists, but will also eventually lead to other courses and programs that can open up the world of energy engineering to data scientists in other fields looking to make a meaningful contribution in the fight against climate change.

Indeed, our experiences with designing and delivering this course indicate that students in graduate energy programs across Europe are highly motivated to learn about data science. Students finishing the course had a solid understanding of how to create their own energy forecasts and use them in decision-making processes. This knowledge can then be used as the foundation for students to take more specialized or theoretical courses, perhaps even in other departments. The three iterations also showed that students, even those without a programming background, could quickly get up to speed with the work required of them while using notebooks. A balance between asking students to understand existing code and writing new code has to be struck here, which we achieved with class projects and competitions.

Going forward, despite their importance, we see such cross-university endeavors to encounter some roadblocks caused by institutional inertia and the fact that graduate students usually come from diverse backgrounds; for example, some engineering programs emphasize programming skills, others do not. From a logistical perspective, keeping curricula up to date and harmonized across universities on an ongoing basis is likewise a challenging issue that we intend to continue exploring in the future. This is especially true for dynamic and rapidly evolving fields such as data science. Common GitHub repositories and shared communication channels, for example, on Slack, have the potential to bridge these intra- and inter-university divides, as well as foster broader communal collaboration.

Disclosure Statement

This work was partly funded by EIT InnoEnergy through a cross-KIC engagement grant on data science for energy. Hussain Kazmi also acknowledges support from Research Foundation–Flanders (FWO), Belgium (research fellowship 1262921N), in the preparation of this manuscript.

References

- Adhikari, A. D. (2021). Interleaving computational and inferential thinking: Data science for undergraduates at Berkeley. *arXiv*. <https://doi.org/10.48550/arXiv.2102.09391>
- Blonsky, M. N. (2019). Potential impacts of transportation and building electrification on the grid: A review of electrification projections and their effects on grid infrastructure, operation, and planning. *Current Sustainable/Renewable Energy Reports*, 6(4), 169–176. <https://doi.org/10.1007/s40518-019-00140-5>
- Brown, J. (2012). The current status of STEM education research. *Journal of STEM Education: Innovations and Research*, 13(5), 7–11. <https://www.jstem.org/jstem/index.php/JSTEM/article/download/1652/1490>
- Bybee, R. W. (2010). What is STEM education? *Science*, 329(5995), 996. <https://doi.org/10.1126/science.1194998>

Edenhofer, O. P.-M. (2011). *IPCC special report on renewable energy sources and climate change mitigation*. Working Group III of the Intergovernmental Panel. <https://www.ipcc.ch/report/renewable-energy-sources-and-climate-change-mitigation/>

Elia. (2021). Adequacy and flexibility study in Belgium: Biennial report.

Feblowitz, J. (2013). Analytics in oil and gas: The big deal about big data. *SPE Digital Energy Conference* (pp. SPE-163717-MS). OnePetro.

García Martín, E. (2017). Energy efficiency in machine learning: A position paper. In N. Lavesson (Ed.), *30th Annual Workshop of the Swedish Artificial Intelligence Society SAIS, Karlskrona (Vol. 137)* (pp. 68–72). Linköping University Electronic Press. <https://ep.liu.se/ecp/137/008/ecp17137008.pdf>

Hong, T. G. (2018). Training energy data scientists: Universities and industry need to work together to bridge the talent gap. *IEEE Power and Energy Magazine*, 16(3), 66–73. <https://doi.org/10.1109/MPE.2018.2798759>

Kazmi, H. & Driesen H. (2020). Automated demand side management in buildings. In M. Sayed-Mouchaweh (Ed.), *Artificial intelligence techniques for a scalable energy transition* (pp. 45–76). Springer. https://doi.org/10.1007/978-3-030-42726-9_3

Kazmi, H. M.-C. (2021). Towards data-driven energy communities: A review of open-source datasets, models and tools. *Renewable and Sustainable Energy Reviews*, 148, Article 111290. <https://doi.org/10.1016/j.rser.2021.111290>

Labeeuw, W. & Deconinck, G. (2013). Residential electrical load model based on mixture model clustering and Markov models. *IEEE Transactions on Industrial Informatics*, 9(3), 1561–1569. <https://doi.org/10.1109/TII.2013.2240309>

Lue, R. A. (2019). Data science as a foundation for inclusive learning. *Harvard Data Science Review*, 1(2). <https://doi.org/10.1162/99608f92.c9267215>

Martín-Páez, T. A.-P.-G. (2019). What are we talking about when we talk about STEM education? A review of literature. *Science Education*, 103(4), 799–822. <https://doi.org/10.1002/sce.21522>

Pathak, B. K. (2016). Emerging online educational models and the transformation of traditional universities. *Electronic Markets*, 26(4), 315–321. <http://doi.org/10.1007/s12525-016-0223-4>

Sokołowski, M. M. (2020). Renewable and citizen energy communities in the European Union: How (not) to regulate community energy in national laws and policies. *Journal of Energy & Natural Resources Law*, 38(3), 289–304. <https://doi.org/10.1080/02646811.2020.1759247>

Strubell, E. G. (2019). Energy and policy considerations for deep learning in NLP. In *57th Annual Meeting of the Association for Computational Linguistics*. Retrieved from <https://doi.org/10.48550/arXiv.1906.02243>

Sweeney, C. B. (2020). The future of forecasting for renewable energy. *Wiley Interdisciplinary Reviews: Energy and Environment*, 9(2), Article e365. <https://doi.org/10.1002/wene.365>

Uddin, M. & Rahman, A. A. (2011). Techniques to implement in green data centres to achieve energy efficiency and reduce global warming effects. *International Journal of Global Warming*, 3(4), 372–389. <http://doi.org/10.1504/IJGW.2011.044400>

van Ruijven, B. J. (2019). Amplification of future energy demand growth due to climate change. *Nature Communications*, 10(1), Article 2762. <https://doi.org/10.1038/s41467-019-10399-3>

Zhongming, Z. L. (2021). *AR6 climate change 2021: The physical science basis*. <https://www.ipcc.ch/report/sixth-assessment-report-working-group-i/>

©2022 Hussain Kazmi, Íngrid Munné-Collado, Khaoula Tidriri, Lars Nordström, Frank Gielen, and Johan Driesen. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.