Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts

Editors:

Biehler, Rolf
Budde, Lea
Frischemeier, Daniel
Heinemann, Birte
Podworny, Susanne
Schulte, Carsten
Wassong, Thomas

Paderborn University

How to cite this publication:
Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts*. Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

Contents

PREFACE

Data science is increasingly relevant in more and more areas of everyday life - but the general education at school so far has hardly responded to these specific changes in digitalization. Completely new challenges for the teaching of mathematics and computer science have emerged, as well as for the subjects of the social and cultural sciences field and for cross-curricular media education. All these subjects have to be reinterpreted in regard to the raising attention to data science, to big data and in regard to a fundamentally changing world of labor and economy. As a reaction, schools have to realize a broad general education that is to be newly defined on the one hand. On the other hand, school education has to stimulate and promote interest in the current, exciting and dynamic new scientific area of data science with its numerous applications.

This book presents the extended abstracts of the presentations at a symposium held in November 2017, discussing economic, social and cultural impacts of big data and data science with experts in curriculum development and educational research in statistics and computer science as well as experts from different facets of data science and its applications. Moreover, analyses from a socio-cultural perspective were included. The main goal was to inspire ideas for teaching data science in secondary schools.

The papers in this volume present perspectives for data science education at school level, collecting contributions from different perspectives: statistics, computer science, sociocultural studies, as well as from institutions, media, and companies.

As a special element in the symposium, we have asked five participating discussants to observe the discussions and focus on five key dimensions of a core curriculum for data science: rationale, aims and objectives, content, tools, best practice. The five discussants presented their view on the presentations and discussions at the symposium in the final panel and summarized their views in this volume.

The symposium was organized by an interdisciplinary team from statistics education and computer science education from the University of Paderborn, working on a project that aims at developing innovative curricula for data science for upper secondary schools. It was initiated and funded by Deutsche Telekom Stiftung (https://www.telekom-stiftung.de/en) and supported by the German Centre for Mathematics Teacher Education DZLM (http://www.dzlm.de/dzlm/international-visitors). As a result of the symposium, the Deutsche Telekom Stiftung has given a grant to the University of Paderborn for a pilot project on data science at school level. The project (http://www.prodabi.de) will develop and test building blocks for curricula at upper secondary level. The project has started in May 2018, the teaching experiment starts with 18 secondary students in September 2018.

We are grateful for the lively and illuminating discussions at the symposium. We want to thank all participants for their valuable input, their engagement and inspiration for the project and to the five discussants for reflecting and summarizing the contributions. A special thanks goes to the Deutsche Telekom Stiftung for their financial support as well as to Ekkehard Winter and Gerd Hanekamp from the Stiftung for their additional support and encouragement. We are very grateful to Tim Erickson who made a careful check of all contributions from a native speaker perspective and from the perspective of data science education. We hope the reader will also get a glimpse of these inspiring days when reading the papers.

Paderborn, September 2018

Rolf Biehler, Lea Budde, Daniel Frischemeier, Birte Heinemann
Susanne Podworny, Carsten Schulte, Thomas Wassong

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

1 Overview

PERSPECTIVES FOR AN INTERDISCIPLINARY DATA SCIENCE CURRICULUM IN GERMAN SECONDARY SCHOOLS

Rolf Biehler & Carsten Schulte
Universität Paderborn, Germany
biehler@math.upb.de, carsten.schulte@uni-paderborn.de

We present some initial guidelines and ideas for an interdisciplinary data science curriculum in German secondary schools, based on a brief discussion of educational philosophy, as well as thematically relevant approaches and traditions in teaching and learning mathematics and computer science.

INTRODUCTION

Data science and its associated buzzwords, for example *big data*, are more and more seen as relevant for education, but so far, few attempts have been made to introduce the field at the school level. Data science is not yet part of the curriculum in secondary schools in Germany and most other countries.

In this introductory chapter, we discuss the goals of the symposium to inform future curriculum development, and we analyze experiences and curricular traditions from the two main subjects, computer science education and statistics education, on which we can build and that we relate to each other. Moreover, we will discuss principles of curriculum development on which we will base our approach.

We also present a first look into our current idea of a future data science curriculum for secondary schools. As a step towards helping teachers and schools implement aspects of data science, we plan a collaboration with selected schools and teachers to implement and try out some ideas of data science in actual secondary classrooms. We subsequently plan to implement a year-long course of 3 hours per week for students who volunteer to take the course, which will in turn become an elective duty course for them that counts for the final examination. We will test ideas from data science with students, establish relations with teachers who are co-designers, and develop material for classroom teaching and professional development courses. Based on that work, we will work on a position paper describing essential components of data science for secondary students.

CHALLENGES FOR DESIGNING A DATA SCIENCE CURRICULUM FOR SCHOOLS

The first goal of the symposium is to exchange ideas, material, and experiences of ongoing projects on data science at school level, at tertiary level, and in internal training programs in companies. As a second goal, an understanding of "fundamental ideas" of data science should emerge: views of data science as a scientific discipline including its relation to statistics and computer science and its historical development, current state, and future perspectives. The notion of fundamental ideas has successfully been used in statistics education to orient curricular developments and teacher actions in the classroom (Biehler 2014a, 2014b; Burrill & Biehler 2011). Thirdly, we want to identify relevant and typical applications of data science and uses of big data in economy, industry, and society; consider ethical, political, legal, and social responsibility aspects of these applications; and reflect on their educational relevance and potential for being made accessible to school teaching. We view our future students in two different roles: working as data scientists themselves or exploring existent systems with the aim of developing simple system models; and also making the black boxes more transparent and identifying underlying assumptions, including economic, political, and cultural conditions and interests. Last but not least, societal and cultural conditions and implications have to be discussed. These aspects are already being analyzed in media education research, in digital humanities, in socio-informatics, and in many popular books (Aoun & ProQuest (Firm) 2017; Harari 2017; O´Neil 2016; Spitz 2017; Weigend 2017). For a data science curriculum, the question is how to educate students so that they can take a thoughtful position in these debates, and on a more practical level, how to integrate societal issues with formal and technical aspects of data science as scientific discipline. The societal aspects include questions about providing private data to companies, critical media competence, "News competence" (dealing with "fake

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

news"), statistical literacy (media reports including products from data journalism and scientific studies using data), and using data and/or data science for one's own goals and in everyday situations.

In sum, various aspects have to be integrated into an overall educational philosophy that comprises all the areas we mentioned; see Figure 1.

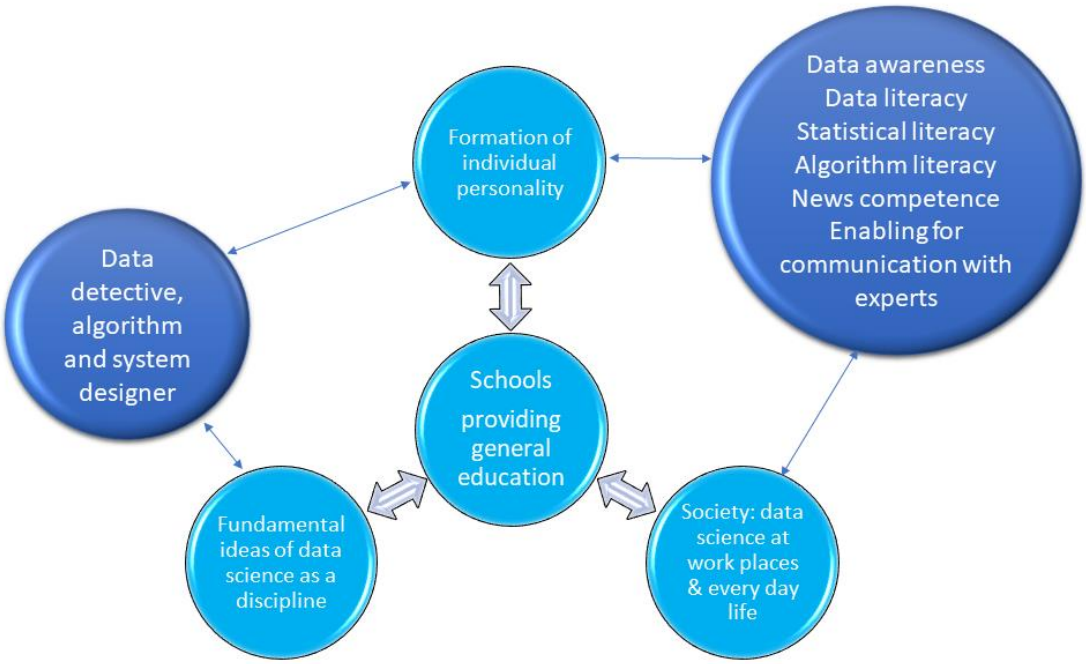


Figure 1: Facets of an educational philosophy for data science

When thinking about curricular reform for school, the question of the German tradition of *Allgemeinbildung* e.g., as expressed by W. Klafki (1996) emerges: Why teach, for whom, with what goals? The underlying so-called "rationale" of the curriculum addresses these questions.

In general, the overarching goal should adhere to self-determination, responsible actions, developing interests, and being introduced to basic ideas of the discipline. Within the context of our data science curriculum development project, we agreed upon four basic guidelines for the curriculum:

1. Develop practical educational resources
2. Figure out and teach fundamental ideas of data science
3. Ensure practical relevancy for everyday life by
    a. identifying relevant application areas
    b. reflecting whether this is of educational value for students
4. Integrate societal and cultural aspects of data science

FRAMEWORK FOR CURRICULUM DEVELOPMENT

Developing a curriculum can be difficult, because a variety of levels, aspects, and people have to be involved.

It can be surprisingly challenging to define the term *curriculum*. Thijs & van den Akker (2009) suggest a broad definition as a "plan for teaching" that can be observed or represented in different levels for various stakeholders. Usually, as in this project, the curriculum is presented as a written document describing an idealized plan for teaching. This plan serves teachers and schools as a point of reference to implement data science in their local classrooms, and probably to derive their own local school-wide curriculum model.

As a plan for teaching, a curriculum model describes several aspects, e.g.: the goals of the teaching, the content, some teaching methods, maybe some specific examples and materials, guidelines for assessment, and so forth. In the curriculum model from the SLO (Netherlands Institute

for Curriculum Development), these different aspects are presented as a curricular spider web (Figure 2).

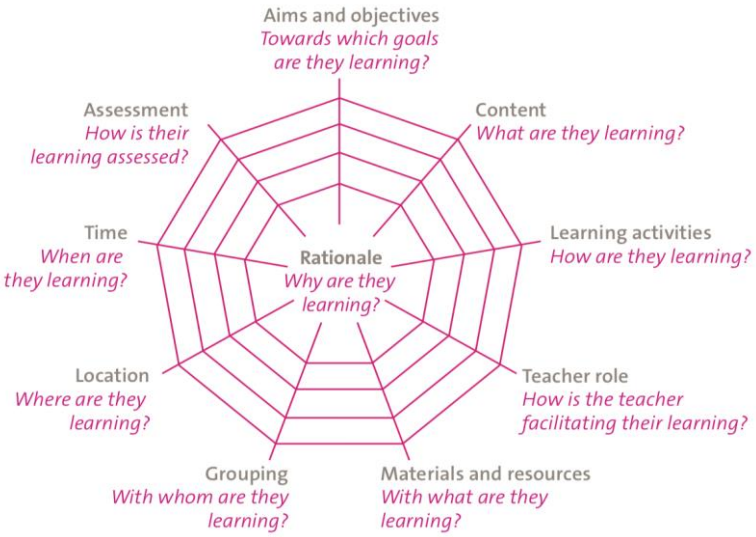


Figure 2: The curricular spider web, taken from Thijs and van den Akker (2009, p. 59)

The figure presents the different dimensions of a curriculum as a spider web to highlight the notion of interdependent and mutually connected aspects that have to be coherent in order to form a suitable curriculum model—or the web will rip apart. Secondly, the graphical presentation highlights the need for an underlying rationale: a philosophy and maybe implicit understanding behind the dimensions that ensures such coherence.

On the level of a teacher, this rationale can also be seen as shared understanding or belief in the nature of the discipline, the core aspects and goals of the subject. The data science symposium and papers in this publication can be interpreted as an attempt to develop such a shared understanding, by inviting experts from different subjects and contexts, and to watch for commonalities, especially in the implicit understanding of the "nature of data science." We thus included experts as observers who presented their view on shared themes as well as differences in the final panel.

The presentations in the symposium focused on perspectives on data science from the academy from business, and from international schools. In this paper, we highlight some important aspects of the curriculum based on curricular traditions in German schools for the two closest subjects to data science, namely statistics education and computer science education.

DATA SCIENCE EDUCATION FROM THE PERSPECTIVE OF STATISTICS EDUCATION

We see the following dimensions for rethinking important impacts of statistics education

- Work flow: Updating the PPDAC cycle
- Extending the statistical view of "data"
- Taking into account Extended and new methods for data science
- Selecting digital tools for data science that support data analysis, data management, and algorithm design
- Taking into account important insights from the statistical literacy discussion

A recent paper that discusses consequences of the data revolution to statistics education is Ridgway (2015).

Work flow

Many statistics educators base their view on the process of statistical inquiry on the so-called PPDAC cycle (Figure 3).

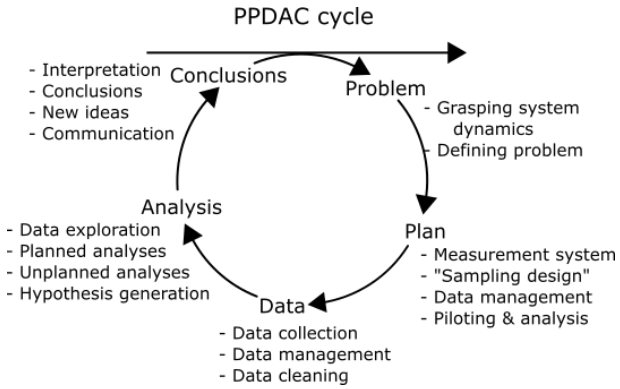


Figure 3: PPDAC cycle according to Wild and Pfannkuch (1999, p. 226)

In books on data science different work flow diagrams are used. We quote from Berthold, Borgelt, Höppner, and Klawonn (2010).

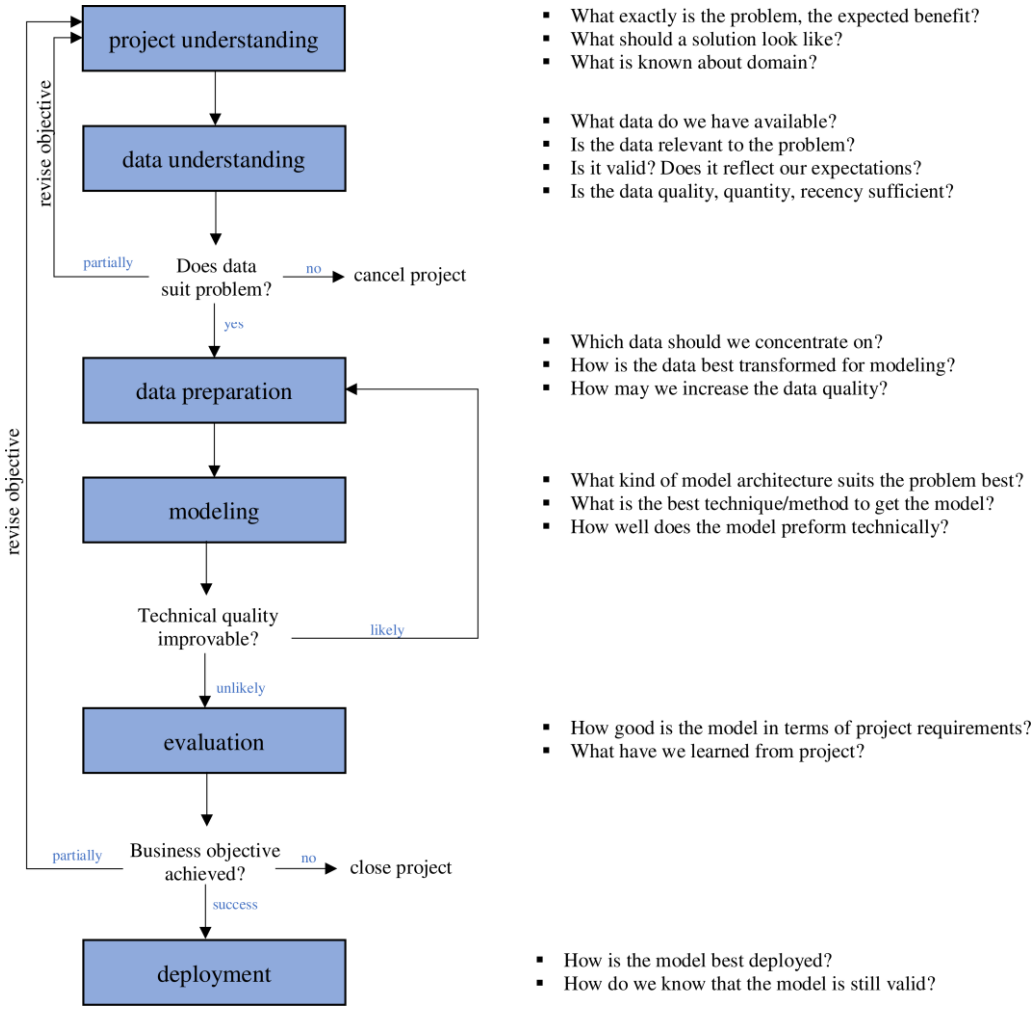


Figure 4: Data analysis cycle according to Berthold et al. (2010, p.9)

We notice several differences between the two cycles; some of them are important for future data science processes.

- Data may be "already there" and not collected according to a plan; the starting point of the cycle may be therefore different (Huber 2011; Tukey 1962)
- Data preparation and data cleaning are regarded as much more important than in statistics
- Project and data understanding are emphasized as part of the "problem" step
- Modeling as a step should be added to the PPDAC cycle
    - Classical statistical theory assumes the model to be given
    - Data science uses new types of algorithmic models (Breiman 2001)
    - Validation of the model is missing as a step (cross validation; distinguishing data for training and for testing is essential)
    - Prediction as a goal for modeling has to be emphasized
- "Conclusions" as a final process step has to be extended
    - Statistics aims at "knowledge"; but data science and computer science "deploy" models. This includes social responsibility, an important step

Extending the statistical view of "data"

The following aspects will be new given the current minor role of data in the school curriculum, where univariate numeric data dominate the statistics curriculum, and bi- and multivariate data are rarely curricular topics.

- Standard in statistics but not in school:
    - The rectangular data tables with different variable types
    - Data of moderate size, multivariate data
- New types of data for statistics
    - Data collected by sensors
    - Data collected by personal devices
    - Transactional data (traffic, supermarket buys)
    - Images and texts
    - Data scraped from webpages
    - Data with geographic information
- Big data; open data

Traditional statistics education often focusses on data with high quality, stemming from controlled randomized experiments or random samples of a well-defined population. Exploratory data analysis in the tradition of John Tukey has always been open to "dirty data," while remaining aware that the kind of conclusion one can draw depends on the quality of the data and a deep knowledge of meta-data. This problem is exacerbated by the many available open and big data sets on the internet, whose origin and quality is often unclear.

Extended and new methods for data science

From the perspective of statistics education, methods such as machine learning, algorithmic models, decision trees, and clustering are new to the curriculum and not yet accessible to school students. We have to cope with the situation that sometimes methods known in statistics get a different name, such as *regression*, which has become one of the methods of *supervised learning*. However, this is not only a new name but a different perspective, with different uses and different generalizations of this traditional statistical method. It may also be the case that new methods can only be introduced in school on the basis of old methods in order to secure better understanding. For instance, one may have to start with bivariate linear regression as a starting point for more complex multivariate and non-linear methods. Teaching bivariate methods from a data science perspective could mean:

- Model fitting with different function classes

- Discussing algorithms for fitting, not treating them as black boxes
- Dealing with model selection, overfitting, and cross validation
- Using different "score functions" (not only least squares)
- Emphasizing validation (residual analysis)
- Using nonlinear regression (smoothing)
- Being aware of different study goals: Explanation or accurate prediction

Selecting digital tools for data science

Currently, in German schools, not much technology is used to support probability and statistics education. If we see technology at all, we see uses of spreadsheets, Geogebra, and graphic calculators, but no statistics tools. Only in experimental classrooms, tools especially designed for supporting the learning and doing statistics and probability are used, such as Fathom and Tinkerplots (https://www.stochastik-interaktiv.de, https://www.tinkerplots.com, https://fathom.concord.org), and building on Tinkerplots and Fathom, the data exploration environment CODAP (http://codap.concord.org, see also the contribution of Bill Finzer to this volume). The question of requirements for digital tools in statistics education (Biehler 1997; Biehler, Ben-Zvi, Bakker, & Makar 2013) has but recently broadened the view by including requirements from data science (McNamara 2015). Whereas the data science at school project led by Rob Gould (see his contribution to this volume) has decided to use R with an adapted set of commands, McNamara also includes Jupyter Notebooks in her review, which can be used as an environment for Python, making it an advanced, relatively easy-to-enter-into programming environment for computer scientists. A growing number of books covering data science with Python have been published (Grus 2016; Haslwanter 2016; Igual & Seguí 2017; McKinney 2017). It is an open question which tools can be adequately introduced at what age level, and whether one should aim at one single tool or use an "entrance tool" for easy data exploration such as CODAP for easy data exploration and then move on to more advanced tools such as R and Python. The latter would, in any case, require the compilation of a student-oriented library of commands, algorithms, and programs. Using Jupyter notebooks (Toomey 2017) may be supportive of this endeavor, but at its root, a curriculum has to incorporate strategies to support an adequate "instrumental genesis" for this these tools for working on data science problems (Guin & Trouche 1999; Madden 2013).

Insights from the statistical literacy discussion

Last but not least, designing a data science curriculum could profit from the lessons learned from the statistical literacy discussion, which takes cultural and societal aspects into account as well as education for critical thinking (Gal 2002, 2003).

We see the following facets:

- Problems of measurement (operationalization of variables, adequacy problem)
- Biases in sampling
- Distinguishing observational studies from experimental studies
- Random assignment and the problem of confounding variables
- Simpson's paradox; ecological fallacy
- Confounding of conditional probabilities
- Understanding visualizations of complex data (including interactive ones)

Various new perspectives that already take data science aspects into account have recently been published (Gould 2017; Grant 2017; Prodromou & Dunne 2017; Schield 2017; Sutherland & Ridgway 2017).

DATA SCIENCE EDUCATION FROM THE PERSPECTIVE OF COMPUTER SCIENCE EDUCATION

In this section we focus on the following aspects, which we will discuss in the next paragraphs:

1. Updated view of data
2. A model of the data science process
3. Integrating societal aspects
4. Tools and resources

Updated view of data

In computer science at school a certain understanding (probably implicit) of data is already taught—and probably needs to be updated or adapted in light of data science.

This model is depicted in a widely accepted framework for educational standards for computer science in Germany, developed by the German *Gesellschaft für Informatik* (GI; see Brinda, Puhlmann, & Schulte 2009), in which a certain understanding of data is taught in computer science at school. The core aim is to emphasize the difference between data and information: Information is construed as understanding data, which happens only in a human mind. The term data is used to label the representation of data in a machine. In this model, a computational device can process, represent, and visualize only data, but not information (see Figure 5).

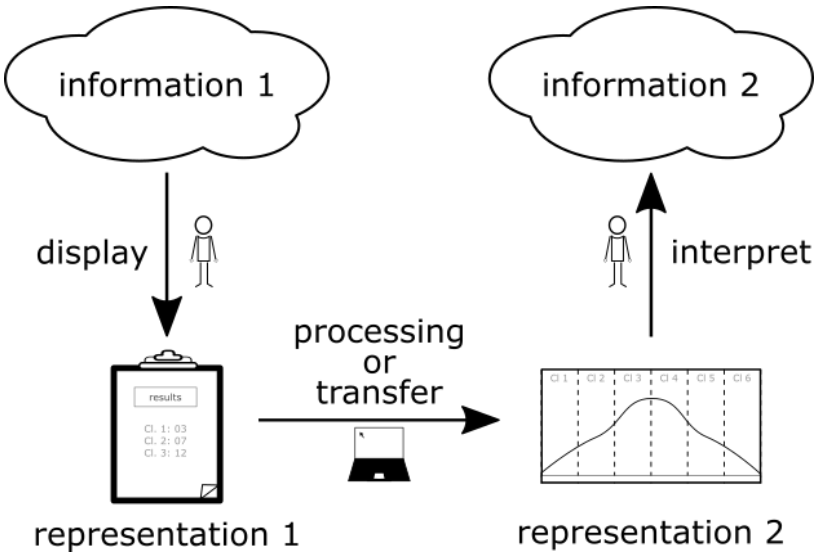


Figure 5: The GI-model of Data vs. Information

Thus, the role of computing technologies should be made clear: we see computing as syntactical operations on data, driven by algorithms working on data represented in suitable data structures. In lower secondary education, the focus is more on operations in standard applications, as depicted in Figure 5, where a data table is (e.g., by an appropriate spreadsheet visualization) transformed into a line graph. In connection with such transformations, rules for appropriate visualizations should be learned, allowing for meaningful interpretation.

In upper secondary education, the focus is more on modeling suitable data structures to store, organize, and retrieve data. Here the focus is on databases and SQL. However, the current focus probably needs to be adapted to incorporate other forms of "data management" in the curriculum (Grillenberger 2014).

The model of data vs. information serves well the intended purpose of highlighting the difference between human and technological data/information processing. In data science, however, more differentiated views on data are used, although there is no consensus on one core model. A popular one, the DIKW-pyramid, distinguishes between **d**ata, **i**nformation, **k**nowledge, and **w**isdom; in connection with this model there is a discussion on the problems of defining knowledge (in contrast to information) as an exclusively cognitive phenomenon (see Wikipedia 2017). In our data science curriculum, we probably need to develop an appropriate definition and educational model of "data" and "information," with regard to the open question whether some levels of understanding (== (information)) should be construed as an exclusively cognitive capacity of the human mind that

eludes computational data processing. This is especially important with regard to artificial intelligence and how this will and should be included in the data science curriculum model.

A model of the data science process

The above outlined model of data vs. information is also based on a classical idea of computing and the general structure of algorithmic processing: First, data is represented (Input); then algorithmically processed or computed (Processing); and then the result is presented via a user interface (Output). This IPO-model is in line with classic algorithmic problem solving—and also echoes the underlying educational model of problem solving and computational thinking as an important learning goal. In this view, students learn to analyze a problem, design a solution, and then implement and test the solution. The solution is the program (The P in the IPO model). The roles of the human and the technology are strictly separated. It can be doubted whether this model is suitable as general problem solving process (with technology), see e.g., Tedre & Denning (2016). In current interactive systems, the user is not only applying pre-defined solutions but interactively designing such solutions. Hence, the strict separation between tool-building and tool-usage becomes blurred.

In terms of the IPO-model, this can be seen as a quick succession of IPO-cycles with immediate feedback. For example, using a standard tool, one can quickly produce different types of data visualization, tweak and adapt them, and decide which general type of visualization (e.g., bar plot vs. pie chart) to use, based on the result. So overall, the problem-solving process is not done prior to technology use, but interactively while using computing. Such a problem solving process is not captured by the idea of structuring data and designing and implementing algorithms, but needs to take into account the socio-technical system or hybrid system in which the human operates (Schulte, Sentance, & Barendsen 2018).

Overall, when problem solving in data science relies on computational tools, the question arises: to what extent do they need to be transparent, and to what extent can they be treated as black boxes? When thinking of using a pre-defined tool to transform some data in a basic visualization (pie chart, bar chart, line chart) it would seem that using the tool as a black box is OK. But what if machine learning is used in a data science course: Would it be appropriate to treat it as a black box, too? Probably not.

Similar questions arise with regard to the aims of project-based learning and the roles of tools: Is it about implementing, or using, or understanding? Is there a need for a new educational approach for the role and/or a new role of problem solving (Tedre & Denning 2016)? And, in summary, what is the need for a new educational process model for data science projects at school?

Data processing can be seen as a central notion in a data science curriculum, while traditional models for projects in computing education focus on software or programming projects. The question is whether or how far these models need to be adapted to data science projects. In computer science education at school, models for organizing software projects are common. Such a process model resembles models from professional software engineering, so that students in class can learn from experiencing a software project. As data science also strongly focusses on "data projects," we will discuss some lessons learned from software projects in education.

Originally, educational projects were oriented on the traditional "waterfall" model (Royce 1970), but subsequently more and more cyclic approaches and elements from agile process models were included. Together with this methodological shift, the learning goals changed, too. Originally the idea was to enable hands-on experiences in an authentic manner, but the increasing size and complexity of real software projects revealed that these expectations were unrealistic. In response, expectations shifted towards more general goals such as learning teamwork, problem solving, and only on some aspects of software development—especially those connected to earlier phases and less with constructing real production systems. The earlier phases focus more on developing ideas for solutions. This development leads to more focus on modeling, and less focus on programming.

Sometimes a consequence is that a successful project probably doesn't have to work, but it suffices to demonstrate (only) a promising idea for solution. See Berger (2001, p 277ff) for a discussion of teachers' expectations. Berger interviewed teachers who teach both math and computer science; this group is likely similar to our prospective data science teachers. In his study he concludes that teachers are likely to be satisfied with promising ideas as the result of a project, and regard problems with implementing the solutions as more or less irrelevant. With regard to software projects

this means that (only) prototypes are designed and built; nevertheless, a prototype necessarily shows some characteristics of a real solution. We are not sure how this will be with data science projects: Does a preliminary and in some aspects wrong data analysis help the student understand the real meaning of the data, or will it lead to wrong interpretations of the data, and in turn also to a wrong understanding of data science and data science methods?

It might be that teachers (and students alike) who are used to seeing projects as developing prototypes tend to trust their preliminary data analysis too soon and forget to systematically rule out counter explanations. In summary, a suitable and useful conceptualization of data projects is a crucial question for the data science curriculum.

There are yet more issues to projects:

First, differences between industry and education: In computer science education it became clear that the overall goals for projects in industry versus education are different (Schubert & Schwill 2011): In industry a group of highly skilled and trained experts works together in a team to produce a working solution; in education a group of untrained learners work together to learn something together. In the educational context, therefore, the division of tasks and e.g., the forming of sub-groups has to be done in a way that ensures the same learning opportunities for all.

Second, from the perspective of computer science curricula, data science projects add a new approach to problem solving. Classically, the process is roughly organized in phases like analysis of the problem, designing a solution, and implementing the solution (with probably several iterative and cyclic steps added). When machine learning on large data sets is applied, then the solution is not directly designed by a human, but "learned" by the machine, based on training data.

While there is this new approach, pragmatically, the overall approach to machine learning is still chosen or influenced by humans, but probably on another level. Such human influence or participation may take place in each of the phases of a data process.

We thus believe it makes sense to conceptualize data science education and data projects in the context of "hybrid" systems. In a report on the future of jobs, the consulting company Cognizant formulated the idea of a hybrid system in terms of a future job, named "Man-Machine Teaming Manager," whose task is to "help combine the strengths of robots/AI software (accuracy, endurance, computation, speed, etc.) with the strengths of humans (cognition, judgment, empathy, versatility, etc.) in a joint environment for common business goals. […] The end goal is to create augmented hybrid teams that generate better business outcomes through human-machine collaboration." (Pring, Brown, Davis, Bahl, & Cook 2017, p. 30). By replacing the focus on business with a focus on society and societal aspects in general, the impact of this view becomes more apparent for education. In Schulte et al. (2018), a first attempt has been made to further elicit this perspective.

Integrating societal aspects

Societal aspects can be viewed as emerging when data science projects are applied. This is also the traditional idea in computer science education, where societal aspects can be discussed in connection with applying software projects. However, the technical view on designing and implementing a new piece of software often overpowers discussion of societal aspects. In addition, software projects often aren't applied in earnest, but are more or less "toy projects" which do not really raise any societal impact or implication. Therefore, unfortunately in computer science education, discussion of societal aspects is more or less decoupled from the more naturally occurring technical aspects of software projects. Similar issues will probably arise with data science education and data science projects. We hope that the notion of hybrid systems helps to integrate societal aspects.

Another way to support teachers and the implementation of the curriculum is to help them to teach ethical issues through carefully-designed curricular material / teaching examples. One specific question that arises for the data science curriculum is whether it makes more sense to integrate reflection on societal aspects into data projects, or instead to establish a learning phase or learning module that solely focuses on societal issues. The first version probably makes inclusion of societal aspects more natural for teachers, whereas when presenting such issues in isolation, teachers are probably more inclined to leave out these aspects and think that societal issues should be taught in social science subjects at school.

With regard to curriculum emphasis probably teachers differ in their approach to inclusion of societal aspects. While we do not have empirical data for data science teachers, a study with chemistry teachers showed that societal aspects are regarded as less important by German teachers (in comparison to other curriculum emphases), see (Driel, Bulte, & Verloop 2008; Markic, Eilks, van Driel, & Ralle 2009).

Tools and resources

In order to do data science, computational tools are needed. We can broadly distinguish two types: On the one hand are interactive tools like spreadsheets, which let one directly manipulate and visualize data. On the other hand are tools like RStudio or Jupyter Notebooks, in which data manipulation is done by using programming languages like R or Python. The first type of tools relies on the "What you see is what you get" and "direct manipulation" paradigms, which aim to create the impression that the user directly works with the data and gets direct feedback. Programming on the other hand is more indirect, as first a set of data manipulations is coded in a formal syntax, and then applied. Both approaches have their merits and fallbacks; e.g. WYSIWYG-tools are easier to use, but programming tools are better for checking how data was manipulated: one can change the script and run again on the original data. We think that both types of tools should be introduced and reflected on by the students.

From the computing education perspective, the intention is not only to use tools in order to learn data science (learning with tools), but also to learn about tools. Learning about tools includes understanding the role and influence different tools have on the data science process, and to understand that tools are designed and constructed for a purpose—and hence that tools can be re-designed. Often, easy-to-use interactive tools are not readily open for re-design by a user, whereas the more programming-like tools afford and inspire adaptation to one's own need.

OUR APPROACH TO CURRICULUM DEVELOPMENT

As outlined at the beginning of this chapter, we draw on the model by Thijs & van den Akker (2009). From the different approaches described there we take on a mix of the communicate and the pragmatic approach. In terms of the communicative approach we have organized a data science symposium and invited experts to discuss perspectives for data science at school from differ related perspectives (see the other chapter in this report), and we have help from four observers which at the end of the symposium reflect their impression on the results of the discussion. This is in line with the approach where the aim is to reach a consensus among experts. From this perspective more such discussion would need to take place with drafts of the curriculum. On the symposium no curriculum draft could be designed, but first promising perspectives be discussed.

In terms of the pragmatic approach we will collaborate locally, and implement a draft of the curriculum in one school: the idea is to meet the requirements of the users, the teachers at school, and get frequent feedback on the curriculum draft via its implementation at school. In terms of the pragmatic approach a cyclic development with refinement of drafts would be way to go. See Thijs & van den Akker (2009) p. 19.

According to them, sustainable curriculum development is based on the synergy with teacher development, and school organization development. For the Katter one option is to employ so-called project courses in upper secondary school, which are open for school curriculum development.

For teacher development that means we have to think about teacher education, too.

One way to go—in line with the above outlined general approaches to curriculum development—is to use design-based approaches and educational reconstruction.

In this process, (Thijs & van den Akker 2009) p, 35ff, suggest to "focus on elements that are essential for the innovation and which may, at the same time, be considered vulnerable as a result of possible complexity or lack of clarity."

CONCLUSION

Designing a curriculum for data science is a challenging task due to a number of issues: Its interdisciplinary nature, complex prerequisites, fast developments, broad application areas, and its relevance for future lives of the students, not to mention the missing teacher education in the area of data science.

There are a number of challenges: The likely heterogeneity of students; limited knowledge on students interests and prior knowledge (and misconceptions); the tension between open exploratory and project-based teaching and learning phases and the need for systematic development of competencies; the broad and complex nature of the field and limited experiences in educational reconstruction and reduction of topics for data science at school. Luckily, we can draw on some resources, as discussed above, e.g., the traditions in computing and math education that already include some aspects of data science, the experiences and inputs from the experts in the symposium, and so forth. In addition, we will focus our task on what, in our view, are some of the most important aspects of the curriculum: the rationale, aims and objectives; and the role of tools (and best practice). These dimensions were also the dimensions our panelists suggested we should especially focus on in observing and reflecting on the symposium.

We plan to design the curriculum in the following cyclic steps:

1. In order to make the planned curriculum live, teachers need to implement it in school. We thus aim to develop material for implementing the curriculum (lesson plans, assessment, …) collaboratively with teachers of pilot schools; and to reflect on the outcomes with experts.
2. We plan to enrich the developed material with description of the underlying rationale, pedagogical goals, and teacher guidance.
3. In addition, we want to develop material for teachers' professional development (**CPD**) courses based on 1. and 2.

REFERENCES

Aoun, J., & ProQuest (Firm). (2017). *Robot-proof : higher education in the age of artificial intelligence*. Cambridge, Massachusetts: The MIT Press,.

Berger, P. (2001). *Computer und Weltbild: habitualisierte Konzeptionen von der Welt der Computer*. Springer-Verlag. Retrieved from http://helianth.net/pap/cuw.pdf

Berthold, M. R., Borgelt, C., Höppner, F., & Klawonn, F. (2010). *Guide to intelligent data analysis: How to intelligently make sense of real data*. London: Springer.

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review, 65*(2), 167–189.

Biehler, R. (2014a). Leitidee Daten und Zufall – fundamentale Ideen aus Sicht der Statistik. In H. Linneweber-Lammerskitten (Ed.), *Fachdidaktik Mathematik - Grundbildung und Kompetenzaufbau im Unterricht der Sekundarstufe I und II* (pp. 69–92). Seelze: Klett Kallmeyer.

Biehler, R. (2014b). Leitidee Daten und Zufall – Fundamentale Ideen aus Sicht der angewandten Stochastik. In H. Linneweber-Lammerskitten (Ed.), *Fachdidaktik Mathematik - Grundbildung und Kompetenzaufbau im Unterricht der Sekundarstufe I und II* (pp. Downloadbereich). Seelze: Klett Kallmeyer.

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clements, A. J. Bishop, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third International Handbook of Mathematics Education* (pp. 643–689). New York: Springer.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16* (3), 199–231. doi:10.1214/ss/1009213726

Brinda, T., Puhlmann, H., & Schulte, C. (2009). Bridging ICT and CS: educational standards for computer science in lower secondary education. *ACM SIGCSE Bulletin, 41*(3), 288–292.

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In Batanero, C., Burrill, G., & Reading, C. (Eds.), *Teaching statistics in school mathematics: Challenges for teaching and teacher education — a joint ICMI/IASE study: The 18th ICMI study* (pp. 57–69). Dordrecht: Springer.

van Driel, J. H., Bulte, A. M. W., & Verloop, N. (2008). Using the curriculum emphasis concept to investigate teachers' curricular beliefs in the context of educational reform. *Journal of Curriculum Studies, 40*(1), 107–122. https://doi.org/10.1080/00220270601078259

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1–25. doi:10.1111/j.1751-5823.2002.tb00336.x

Gal, I. (2003). Expanding conceptions of statistical literacy: An analysis of products from statistics agencies. *SERJ, 2*(1), 3–21.
Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal, 16*(1), 22–25.
Grant, R. (2017). Statistical literacy in the data science workplace. *Statistics Education Research Journal, 16*(1), 17–21.
Grillenberger, A. (2014). Big data and data management: A topic for secondary computing education. In *Proceedings of the tenth annual conference on international computing education research* (pp. 147–148). New York, NY, USA: ACM. https://doi.org/10.1145/2632320.2632325
Grus, J. (2016). *Einführung in data science - Grundprinzipien der Datenanalyse mit Python*: O´Reilly.
Guin, D., & Trouche, L. (1999). The complex process of converting tools into mathematical instruments: The case of calculators. *International Journal of Computers for Mathematical Learning, 3*(3), 195–227.
Harari, Y. N. (2017). *Homo Deus: A brief history of tomorrow*. New York: Harper Collins Publishers.
Haslwanter, T. (2016). *An introduction to statistics with Python with applications in the life sciences*. Cham, Switzerland: Springer.
Huber, P. J. (2011). *Data analysis: what can be learned from the past 50 years*. Hoboken, N.J.: Wiley.
Igual, L., & Seguí, S. (2017). *Introduction to data science: A Python approach to concepts, techniques and applications*. Cham, Switzerland: Springer.
Klafki, W. (1996). *Neue Studien zur Bildungstheorie und Didaktik Beitraege zur kritisch-konstruktiven Didaktik*. Weinheim u.a.: Beltz.
Madden, S. (2013). Supporting teachers' instrumental genesis with dynamic mathematical software. In D. Polly (Ed.), *Common core: Mathematics standards and implementing digital technologies* (pp. 295–318). Hershey, USA: Information Science Reference (an imprint of IGI Global).
Markic, S., Eilks, I., van Driel, J., & Ralle, B. (2009). Vorstellungen deutscher Chemielehrkräfte über die Bedeutung und Ausrichtung des Chemielernens. *CHEMKON, 16*(2), 90–95. https://doi.org/10.1002/ckon.200910091
McKinney, W. (2017). *Python for data analysis: Data wrangling with pandas, numPy, and iPython*. USA: O´Reillly.
McNamara, A. (2015). *Bridging the gap between tools for learning and for doing statistics.* (Doctor of Philosophy in Statistics), University of California, Los Angeles. Retrieved from http://escholarship.org/uc/item/1mm9303x
O´Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. New York: Crown.
Pring, B., Brown, R. H., Davis, E., Bahl, M., & Cook, M. (2017). *21 jobs of the future: A guide to getting – and staying – employed over the next 10 years*. Cognizant.
Prodromou, T., & Dunne, T. (2017). Statistical literacy in data revolution era: Building blocks and instructional dilemmas. *Statistics Education Research Journal, 16*(1), 38–43.
Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review, 84*(3), 528–549.
Royce, W. (1970), Managing the development of large software systems. *Proceedings of IEEE WESCON*, 26 (August): 1–9.
Schield, M. (2017). CAISE 2016 promotes statistical literacy. *Statistics Education Research Journal, 16*(1), 50-54.
Schubert, S., & Schwill, A. (2011). *Didaktik der Informatik*. Springer.
Schulte, C., Sentance, S., & Barendsen, E. (2018). Computer science, interaction, and the world. In *Computer science education: Perspectives on teaching and learning in school* (1st ed., pp. 57–74). Bloomsbury Academic.
Spitz, M. (2017). *Daten - das Öl des 21. Jahrhunderts? Nachhaltigkeit im digitalen Zeitalter*. Hamburg: Hoffmann and Campe.
Sutherland, S., & Ridgway, J. (2017). Interactive visualisations and statistical literacy. *Statistics Education Research Journal, 16*(1), 26-30.

Tedre, M., & Denning, P. J. (2016). The long quest for computational thinking. In *Proceedings of the 16th Koli Calling International Conference on Computing Education Research* (pp. 120–129). New York, NY, USA: ACM. https://doi.org/10.1145/2999541.2999542

Thijs, A., & van den Akker, J. (Eds.). (2009). *Curriculum in development*. Niederlande: SLO - Netherlands Institute for Curriculum Development.

Toomey, D. (2017). *Jupyter for data science: Exploratory analysis, statistical modeling, machine learning, and data visualization with Jupyter*. Birmingham: Packt Publishing.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33*, 1–67.

Weigend, A. S. (2017). *Data for the people : how to make our post-privacy economy work for you*. New York: Basic Books.

Wikipedia (2017) DIKW pyramid. (2017, November 18). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=DIKW_pyramid&oldid=810990870

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–265.

2 Data science education at university level

ALGORITHMS, KNOWLEDGE, AND DATA: ON THE EVOLUTION OF INTELLIGENT SYSTEMS DESIGN

Eyke Hüllermeier
Heinz Nixdorf Institute
Department of Computer Science
Paderborn University, Germany
eyke@upb.de

During the past decades, the design of intelligent systems and development of applications in Artificial Intelligence (AI) has been subject to a steady evolution. Most notably, there has been a significant shift from the classical knowledge-based paradigm to a strongly data-driven approach. This shift has been fostered by the recent emergence of data science as a scientific discipline and the success of machine learning (ML) as one of its core methodologies. Elaborating on the evolution of algorithm and intelligent systems design in general, this talk will therefore specifically focus on recent developments in machine learning. Proceeding from the standard algorithmic approach as commonly adopted in computer science, three paradigms will be motivated and explained briefly.

KNOWLEDGE-BASED PROGRAMMING AND EXPERT SYSTEMS

The first paradigm is knowledge-based programming and expert systems, which were quite popular in the 80s and 90s of the last century (Puppe 1993). A key innovation of this approach was a strict separation between a domain-specific knowledge base, responsible for the representation of knowledge and facts relevant for the problem at hand, and a domain-independent inference engine, responsible for the processing of knowledge in a logically sound way. Compared to standard algorithms, this approach helped to significantly facilitate the development and maintenance of intelligent systems. Yet knowledge acquisition remained a major obstacle.

DATA-DRIVEN SYSTEMS DESIGN AND MACHINE LEARNING

While corresponding aspects of knowledge representation and reasoning have dominated research in AI for a long time, problems of automated learning and knowledge acquisition have more and more come to the fore in recent years. There are several reasons for this development, notably the following. First, caused by the awareness of the aforementioned "knowledge acquisition bottleneck" and the experience that a purely knowledge-based approach to systems design is difficult, intricate, and tedious most of the time, there has been a shift in research from modeling to learning and adaptation, i.e., from the knowledge-based to the data-driven design of intelligent systems (Hüllermeier 2015). The latter not only suggests itself in applications where data is readily available, but can sometimes even be essential. In learning on data streams, for example, models are not only constructed once (from a static "batch" of data) but need to be updated continuously in response to data arriving in real time (Angelov et al. 2010), which cannot be accomplished by a human expert. Second, this trend has been further amplified by the great interest that the field of knowledge discovery in databases, and its core methodological components, machine learning and data mining, have attracted in recent years (Bishop 2006). Learning from data and data analytics has become a ubiquitous topic in the era of "big data." In contrast to the common conception, however, model induction is never purely data-driven; instead, some sort of prior (domain) knowledge is always required, too. Ideally, the machine learning approach allows for combining knowledge and data in a flexible manner, and for compensating either of the two by the other.

AUTOMATED MACHINE LEARNING

Data-driven model induction significantly reduces the need for domain knowledge. Even so, the design of effective algorithms still requires a high level of ML expertise. Since end users in application domains are normally lacking this expertise, there is a need for suitable support in terms of tools that are easy to use. Ideally, the induction of models from data, including the data preprocessing, the choice of a model class, the training and evaluation of a predictor, the

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

representation and interpretation of results, etc., would be automated to a large extent (cf. Figure 1 for an example of a typical machine learning pipeline). This has triggered the field of automated machine learning (Feurer et al. 2015). AutoML seeks to automatically select, compose, and parametrize machine learning algorithms, so as to achieve optimal performance on a given task.

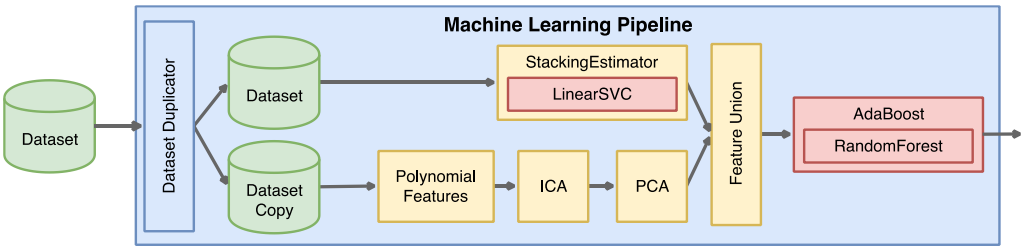


Figure 1: Example of a machine-learning pipeline for extracting a predictive model (the output produced as a result) from a dataset provided as in input. The pipeline includes several steps for preprocessing the data and trains a predictive model (a classifier) in the end.

SOCIAL AND SOCIETAL PERSPECTIVES

With an increasing dissemination of AI technology and usage of algorithms for automated decision making, the discipline of machine learning (and, more broadly, data science) has gained significant social and societal relevance. Many decisions taken by algorithms in real world scenarios, even if unnoticed, have an immediate influence on individuals or groups of individuals. A simple example is search engines on the internet, which decide about the information to be provided or hidden, perhaps on a personalized basis. A more extreme example is the idea of using machine learning to improve jail-or-release decisions of judges (Kleinberg et al. 2017).

In light of applications of that kind, it is hardly surprising that, in addition to more or less objective criteria such as correctness of predictions, other criteria of AI and machine learning algorithms have come to the fore. An important example is the quest for *interpretability* and *explainability*: Algorithmic decisions should be transparent, so that a user can understand and comprehend them. Needless to say, "explainable AI" is especially difficult to realize for data-driven approaches, where decisions strongly depend on the data, perhaps in very subtle ways. Another example is the quest for *fairness*, i.e., algorithms that avoid the (perhaps unintended) discrimination of people (Zliobaite 2017).

CONCLUSION

The increased availability of data in all branches of our daily lives has led to significant advances in the design of intelligent systems: modern AI strongly relies on machine learning methodology to produce such systems in a data-driven manner. These advances on a technical level, which are likely to continue in the foreseeable future, are accompanied with an increased societal relevance and responsibility of AI and data science.

REFERENCES

Angelov, P., Filev, D., & Kasabov, N. (2010). *Evolving intelligent systems*. New York: John Wiley and Sons.

Bishop, C. M. (2006) *Pattern recognition and machine learning*. New York: Springer-Verlag.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. *Proc. NIPS, Advances in Neural Information Processing Systems*, 2962–2970.

Hüllermeier, E. (2015). From knowledge-based to data-driven fuzzy modeling: Development, criticism, and alternative directions. *Informatik Spektrum*, *38*(6), 500–509.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Puppe, F. (1993). *Systematic introduction to expert systems: Knowledge representation and problem solving methods*. Berlin: Springer-Verlag.

Zliobaite, I. (2017). *Fairness-aware machine learning: a perspective*. CoRR, abs/1708.00754, http://arxiv.org/abs/1708.00754

DATA SCIENCE MASTER'S DEGREE

Göran Kauermann
Ludwig-Maximilians-Universität München
goeran.kauermann@lmu.de

With the progress of the information age and the novel challenges in big data analytics in business, industry, science, and society, the demand for trained data scientists is growing rapidly. Many universities have responded to demand and are offering new courses in data science. In this brief note we support these activities and motivate data science as a combination of statistics and computer science. As an example, we present the new master's program data science at LMU Munich.

INTRODUCTION: DATA SCIENCE AS SCIENCE

Data science is not clearly defined as a scientific field. Statisticians often take Cleveland's (2001) seminal and inspiring work as definition, where the title already sets the direction: "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." But the expansion of statistics towards computer science does not completely define the field; the opposite direction is also essential. Cleveland writes: "The benefit to the data analyst has been limited, because the knowledge among computer scientists about how to think of and approach the analysis of data is limited, just as the knowledge of computing environments by statisticians is limited." From our point of view, this combination of statistics and computer science provides the foundation of data science (see also Kauermann & Küchenhoff 2016).

Breiman (2001) categorizes data analysts into two cultures labeled as the "generative modeling" and "predictive modeling" followers. In a simplified form, this can be transferred to the two concrete questions that are pop up in data analysis: "What's going on?" and "What happens next?" While in most cases, the first question can be solved with statistical models, for the second question, machine learning instruments commonly provide better answers. A data scientist, however, must master both perspectives. He or she is in a sense a hybrid, uniting both of Breiman's cultures in one person. Recent definitions of data science are going in a similar direction. Mike Driscoll (Metamarket CEO) is credited with the following quote on the web: "Data Science is a blend of Red-Bull-fuelled hacking and espresso-inspired statistics." We visualize this in Figure 1. Tied together by Big Data issues, statistics and computer science enter the new field of data science.

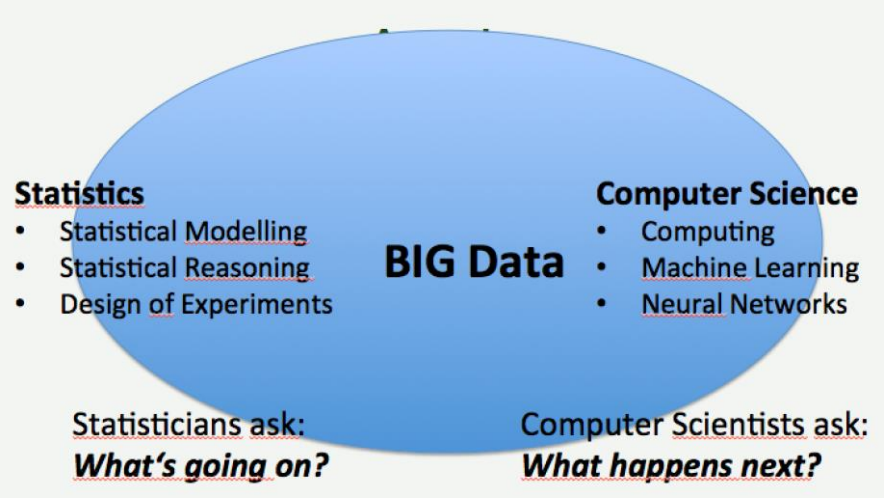


Figure 1: Statistics and Computer Science in Data Science

Data science is most often driven by applications, "use cases" as they are called in the corresponding jargon. In that sense, data science is the intersection of statistics, computer science,

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

and applications. And even though applications are central, it does not mean that data science is purely applying statistics or machine learning instruments. To quote again from Cleveland: "technical areas of data science should be judged by the extent to which they enable the analyst to learn from data. (...) Tools that are used by the data analyst are of direct benefit." In this respect, data science also uses theory and methodology and drives the development of these in the same way. Thus, we can clearly conclude that data science is actually science.

TRAINING: DATA SCIENCE AS MASTER

Data science, as motivated above, is about extracting knowledge and information from data. This requires competencies in statistical data analysis as well as in numerical implementation and data management. Such a mix of skills needs proper training and lecturing and is today offered in a number of new master's programs at universities. Though several new programs in this direction have started recently, not all of them take the mix of statistics and informatics as a starting point. Some are just relabeling either statistics or machine learning as data science, which is certainly an understandable selling point, but not necessarily fair towards the new scientific field.

The Ludwig-Maximilians-Universität München (LMU) has also started a novel program in data science, which was launched winter semester 2016/17. It is one of the first international degree programs in the field of data science offered at a German university (master of science). The data science program is interdisciplinary and is jointly and in equal parts run by the Institute of Statistics and the Institute of Computer Science at LMU, see www.datascience-munich.de. This program is a real joint venture between statistics and computer science. The program is financially supported by the elite network Bavaria (www.elitenetzwerk.bayern.de) and is therefore labeled as an *elite degree program*. The naming also reflects that the program does not primarily want to meet the almost excessive demand for data scientists but aims to train outstanding students. The program is designed for international students and is entirely in English.

The study program is in competition with many newly-created courses, not only in Germany, but also internationally. However, due to the elite status, the degree program can offer special features, making the program attractive. These include summer schools on data ethics, data "hackathons" and/or tutorials on specific topics, and more. In addition to the methodological modules, the study program also includes a practical module to familiarize students with "hands on" real-world challenges. The modules of the degree program are shown in Figure 2. We refer to Kauermann & Seidl (2018) for deeper discussion.

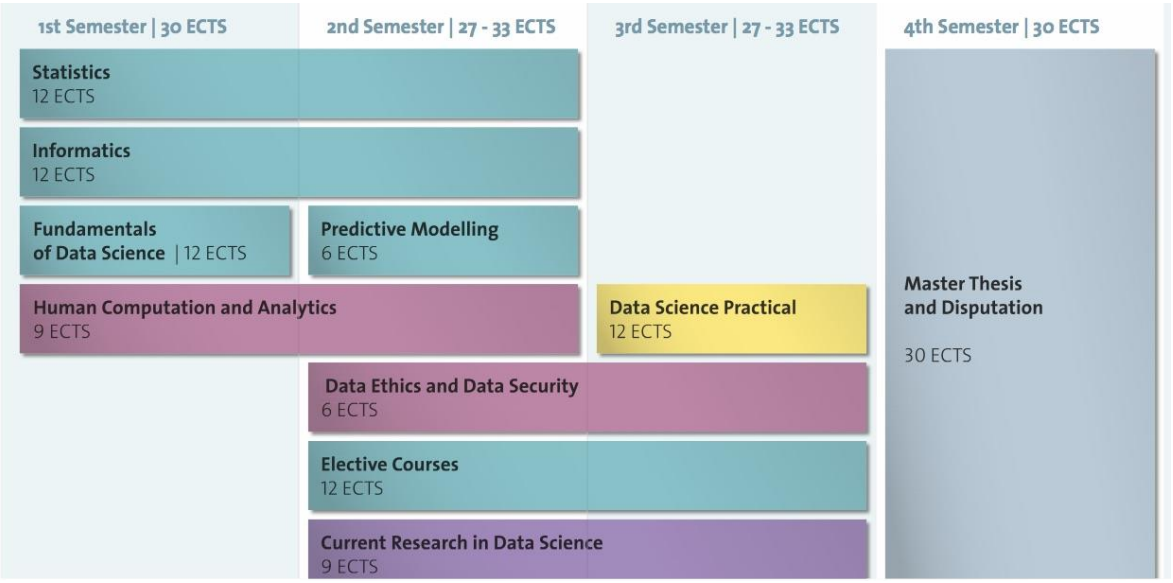


Figure 2: Modules of data science program at LMU

ETHICS: DATA SCIENCE AND DATA ETHICS

Data ethics is a large field and certainly even less clear than data science itself. We consider several aspects under the phrase data ethics. This includes data protection, data security, and legal aspects, as well as ethical views on data usage. These topics exceed the competencies of statisticians and computer scientists, even though technical aspects of data security certainly fall in the field of expertise of computer scientists. However, this is just one side of the coin and legal or ethical views of data usage are not covered in the core disciplines statistics and computer science. Nonetheless, these topics are essential and important and should belong in a data science program as well. At the LMU these fields are covered by external speakers, who are invited to an annual summer school on data ethics. The summer school stimulates attention and problem awareness and demonstrates that close interaction and collaboration between data scientists on the one side and experts in law and/or ethics on the other side is indispensable. This is important, in particular in times where the new European data protection regulation is coming into effect.

CONCLUSION

Data Science is a new scientific field. It opens a new avenue for collaboration between statistics and computer science; and the demand for well-trained data scientists in turn demands new data science master's programs at universities. The syllabus of these programs should mix statistics and computer science. Moreover, data ethics needs to be an essential component of the programs.

REFERENCES

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*, 199–231.

Cleveland, W.S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*, 21–26.

Kauermann, G. & Küchenhoff, H. (2016) Statistik, Data science und big data. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *10*(2), 141–150.

Kauermann, G. & Seidl, T. (2018). Data science: A proposal for a curriculum. *International Journal of Data Science and* Analytics. (to appear). DOI: https://doi.org/10.1007/s41060-018-0113-2

DATA SCIENCES AS ESSENTIAL JOB SKILL! A SOCIAL SCIENCE, ECONOMICS, AND PUBLIC POLICY TRAINING PERSPECTIVE

Frauke Kreuter
University of Maryland, University of Mannheim,
Institute for Employment Research, Germany
frauke.kreuter@uni-mannheim.de

Understanding data, understanding what information can be derived from data, understanding which decisions can be made based on such information, and knowing whom to give data to and when, will be essential for everyone joining the workforce. Even if current students are not seeking degrees in STEM/MINT fields, they will be confronted with data and will be asked to make evidenced-based decisions. This paper describes two continuing education programs that showcase which skill gaps need to be filled and why. The broad outline of these programs can serve as a baseline for discussions on the establishment of data science curricula in high schools. [Note: A portion of this presentation is also given at the ICOTS 2018 and part of their proceedings papers. The work on the International Program in Survey and Data Science https://survey-data-science.net/ is done with Zhenya Samoilova and Florian Keusch; the work on the Coleridge Initiative https://coleridgeinitiative.org is done with Rayid Ghani and Julia Lane.]

MOTIVATION

The volume of data keeps increasing and is expected to rise nearly tenfold by 2025 (Reinsel et al. 2017, p. 3). In the private sector, new types of "big data" are in high demand because data-driven decision-making can help businesses become more efficient and gain substantive insights about their clients. Policymakers hope to improve the viability and effectiveness of government programs and policies by unlocking large quantities of administrative data (Report of the Commission on Evidence Based Policy Making 2017). Statistical agencies are also interested in integrating new digital sources of data into official statistics, as they hope this will allow them to collect more timely data and to reduce response burden (National Academies of Sciences, Engineering, and Medicine 2017). Given the strong interest in data-driven decision-making across various areas, the demand for experts trained to work with different data sources is rising. However, not only is an increased labor force needed to meet this demand, but also, a new set of skills is needed for those already employed and charged with data analysis tasks.

Roberto Rigobon (2015) used the graph below in an expert meeting of the Bureau of Economic Analysis to describe the nature of the change. Based on Groves (2011), he highlights the organic nature of transactional data, or data created with an aspiration in mind (e.g., postings on social media outlets). Unlike data arising from experiments and surveys, data from transactions or internet interactions arise organically, and thus often lack a clear structure or do not fit the intended measurement purposes.

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

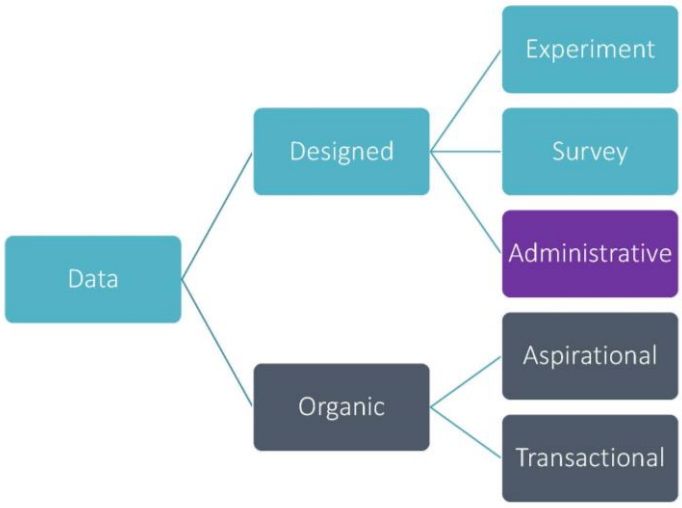


Figure 1 Examples for designed and organic data. Source: Rigobon (2015)

While big data has enormous appeal in its potential to describe economic and social activity, the structure of initial data science training has often been confined to technical training in computer science departments. As a result, while there are many courses in machine learning, database management, or text analysis, the content of those curricula are typically not focused on how to use these skills to contribute to the scientific analysis of social problems. In addition, most such programs are targeted at traditional graduate students, and are part of a continuous course of study. Yet there is a great need for graduate training for nontraditional students in government agencies and the private sector, students who need to understand how to use data science tools as part of their regular employment. The programs presented here are intended both to train a new generation of data scientists and to increase the skill and competency level of the existing workforce.

COLERIDGE INITIATIVE

The U.S. National Science Foundation funded, as part of their Innovate Graduate Education funding pipeline, a consortium from the University of Chicago, University of Maryland, and New York University, to jointly design an executive education program in which working professionals are paired with graduate students to work on research projects (NSF 1633603). After three successful executions of the training program, the Coleridge Initiative was formed (coleridgeinitiative.org).

Academic Overview and Curriculum

The program draws on students from federal, state, and local government agency staff, the private sector, as well as traditional sources. Rather than simply learning data science in the abstract, a major feature of the curriculum is that the program is structured around teams that work on issues of greatest interest to them, and draws on the data, models, and tools best suited to address those issues. The PIs of the project found in the early prototype that this approach has the advantage of pulling together all the main skill sets needed to use big data, as well as training participants in privacy and confidentiality, and producing quite useful insights that can be used in many other policy contexts. The project has several goals for the students:

(i) the development of a broader range of skills applicable in the new economy. In particular, the project will help social scientists make full use of tools used by computer scientists, while enabling computer scientists and statisticians to move beyond the mastery of isolated tools and learn how to apply research methods to a specific problem (for example, scraping data from the web or running a classification algorithm is a tool, but it is of limited use without an understanding of why the data are needed, what process generated the data, and what inference should be drawn from the data);

(ii) the building of a systematic understanding of the steps of problem formulation, deciding what data can be used to solve the problem, devising a plan for collecting that data, processing the data, analyzing it, and finally presenting and publishing the results;
(iii) the acquisition and use of the skills needed for a data-centric research career such as working within collaborative data environments, understanding and mastering data collection and management techniques, and mastering data hygiene, particularly replicable and reproducible methods;
(iv) the acquisition and use of soft skills such as leadership, mentoring, and dealing with clients.

The core of the program is administered in four modules. In the first module, interdisciplinary teams are formed, ensuring a mix of computer and social scientists on each team, as well as a mix of traditional and non-traditional students. Those teams work together throughout the module series. Jupyter notebooks are introduced to serve as learning devices. The module itself covers the fundamentals of problem formulation, inference, and basic coding tools. It ties those fundamentals to a specific topic area. The problem formulation section for the first module includes such topics as: (i) understanding the science of measurement, (ii) identifying research goals, and (iii) identifying measurement concepts and data sources associated with the research goals. The inference section will introduce students to the data generation process, dealing with missing data, and selection issues. The coding section will introduce students to basic Python coding.

After completing the first module, teams work to further develop their research problem before they participate in the second module on "Data capture and curation," which includes web scraping and data capturing through APIs, data linkage, database basics, and programming with big data. Subsequently, teams prepare data for their own research projects, linking if necessary and preparing them for the databases. In the third module on "Modeling and analyses," the focus lies in discussing machine learning techniques, analyses of networks, and text analysis. Teams apply those techniques to their data in the month following this module. In the fourth and final module, students learn about presentation of data, inference, and ethics. During and after this module student work on various visualizations of their data, and structure the core of upcoming presentations and publications. So far more than 40 U.S. agencies from local, state, and federal governments have participated in the program.

INTERNATIONAL PROGRAM IN SURVEY AND DATA SCIENCE

The German Federal Ministry of Education and Research funded the International Program in Survey and Data Science (IPSDS). As newly graduated job candidates often lack sufficient substantive knowledge in their respective work areas, and data training is often sought by people with a strong professional background in various areas (e.g., computer science, finance, consulting, and official statistics), the program focuses on providing training to working professionals who already possess some skills and experience. The program constitutes a consortium of Universities, led by the University of Maryland and the University of Mannheim, and aims to establish an online Master's degree targeting working professionals in the areas of opinion, market, and social surveys, as well as employees of survey research enterprises, ministries, and statistical agencies worldwide. Since entering a traditional graduate program for those who are full-time employed and have family responsibilities tends not to be feasible, an online format for the delivery was essential to making the program a viable and realistic option for the identified target groups. Four-year funding (2014–2017) made it possible to enroll two cohorts of students who participated in the program free of tuition or fees, and in turn helped evaluate and guide the program's development.

Academic Overview and Curriculum

While most data science programs focus their curricula on the areas of statistics and/or computer science, IPSDS adds to these key areas training in data collection and data quality; students are introduced to various types of data, both in isolation and in aggregation. In recent years, it has become clear that new digital sources of data will not replace surveys, but instead will

be used in conjunction with survey data (Japec et al. 2015). Combining both new and traditional data sources is a tenable strategy already applied by leading organizations and businesses. Statistical agencies cannot rely solely on the new data sources, as they cannot capture all relevant aspects of a given research problem. Additionally, the increasing popularity of "Do-it-Yourself" tools, which indirectly indicate the persistent reliance on survey data in the industry, cannot be ignored. For example, Survey Monkey reports 90 million of their online questionnaires are filled out every month, while Qualtrics sends out approximately one billion survey invitations per year (Callegaro & Yang, in press). As the latest report of the Committee of National Statistics of the U.S. National Academy of Science (2017) stresses, using multiple data sources and understanding the data generating process for each data type is of central importance for data quality.

The IPSDS curriculum draws on five areas identified by the task force report of the American Association for Public Opinion Research: (re)defining research question(s), the data generating process, data curation and storage, data analysis, and data output and access (Japec et al. 2015). The program consists of 75 European academic credit (ECTS) points that are equivalent to 15 months of full-time studying (more information about the five core modules and example courses for each of the modules can be found here: https://survey-data-science.net/program/curriculum). Based on the data from the first two student cohorts, part-time studying will take, on average, three years to complete the Masters. All of the students are required to attend the introductory course Fundamentals of Survey and Data Science. Following a successful completion of this foundation course, students can choose from a number of courses within each module. By allowing flexibility, the program can be more responsive to students' career needs, backgrounds, and interests.

To address the needs of the program's target audience, the format has the following features:

- course materials can be accessed asynchronously anytime and anywhere;
- in addition to asynchronous materials access, students attend small live online classes moderated by the course instructor(s), in order to discuss questions and engage in collaborative problem-solving;
- IPSDS runs entirely in English, due to its international focus;
- once a year students are expected to participate in an immersion event (Connect@IPSDS) that takes place at the University of Mannheim, where they can meet their peers and faculty in person, attend talks and workshops by leading experts in the field, network with international survey and data science professionals, and engage in additional learning and community-building activities.

IPSDS online environment and course design

The goal of IPSDS is not only to create courses with high-quality content delivered by leading experts in the field, but to make courses interactive, engaging, academically rigorous, and yet flexible. For this reason, the program adopts the flipped classroom (FC) model. In recent years, the flipped classroom design (also known as "inverted instruction" or "inverted classroom") has received a lot of attention among researchers and practitioners (Prober & Khan 2013). Although there is some disagreement on the exact definition of the FC, literature agrees it includes rotating between two phases: 1) pre-class interaction with course materials and 2) guided learning activities (He et al. 2016). The "flipping" designates the idea of moving what usually was done in the classroom (e.g., lecturing) out of the classroom and bringing traditionally out-of-classroom activities (e.g., work on problem sets or assignments) inside the class.

Pre-class Interaction with the course materials:

The course materials provided to the students on the course website (currently Moodle) include (depending on the course content) pre-recorded video lectures, required and recommended readings, examples for programing exercises, datasets, and other additional resources. Providing students with reading materials prior to the class is not a new phenomenon. In contrast to simply mediating this common practice with technology (upload documents online before class), the FC design stresses the quality of interaction with the provided resources. The integration of pre-recorded lecture videos on the learning platform allows for pausing, moving forward and backward

in the video, and re-watching parts of the video, as well as changing the video speed (both increasing and decreasing the speed is possible). The lectures are broken into several shorter videos to take into account the prototypical attention span as well as relatively short windows of free time among working professionals. Video materials include lectures, interviews, and discussions with experts, as well as demonstrations of specific techniques and software tools.

Guided learning activities:

The in-class guided activities of the FC design are implemented with the help of online video conferencing and discussion forums. Each week, students join a 50-minute online session using Zoom software, moderated by the instructor. The online sessions are mandatory and serve multiple purposes: 1) discuss students' questions, 2) review problems with assignments, and 3) motivate students to persist in the course. Prior to coming to an online session, students are expected to go through the material of the appropriate unit. They are also encouraged to submit their questions to the instructor via the discussion forum or email. The discussion forums are also used to provide opportunities for additional (often optional) communication as well as to refer students to relevant external resources (e.g., news articles, websites).

Each course has its unique virtual room permitting the same link to be used for all of the online meetings. The link is posted on the course website as well as communicated to the students in the welcome email. The moderator (in our case, the instructor) can move back and forth between these rooms. Small class sizes (up to 18 students) make it possible to see every person in the meeting simultaneously. The user can choose to use a speaker view or a full screen view. In the speaker view, the speaker will be shown in the main window (see Figure 2a). In the full screen view, one can see all of the participants at the same time (see Figure 2b).

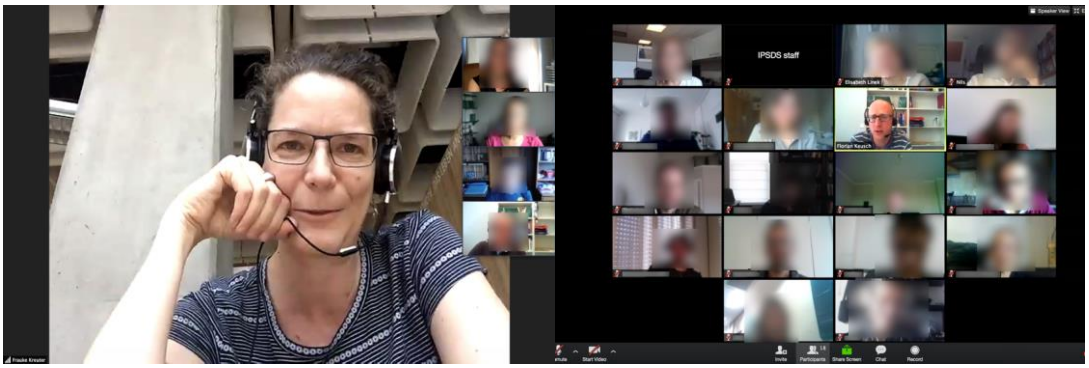


Figure 2a/b (from left to right): a. Zoom interface in speaker view; b. Zoom interface in the full screen view. Faces (other than the authors') are blurred here for privacy reasons.

CONCLUSION

In both programs introduced here, extensive evaluations took place through pre- and post-class surveys, qualitative interviews, and observations, yielding the following insights: 1) the modular approach is much appreciated by working professionals; 2) learning is particularly effective with applications at hand and as part of real-world data analyses projects, ideally embedded into the employers' mission; 3) professionals from all sectors and disciplines are interested in these programs; 4) teaching privacy and confidentiality is very important; and 5) asking the right (research) question is the hardest element to learn and to teach.

REFERENCES

Callegaro, M. & Yang, Y. (2017). The role of surveys in the Area of "Big Data". In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 175–192). New York: Palgrave.

Groves, R. M. (2011). "Designed data" and "organic data." *Directors Blog*, May 31. U.S. Census Bureau. Retrieved January 28, 2015 (http://1.usa.gov/15NDn8w).

He, W., Holton, A., Farkas, G., & Warschauer, M. (2016). The effects of flipped instruction on out-of-class study time, exam performance, and student perceptions. *Learning and Instruction 45*, 61–71.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly 79*(4), 839–880.

National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in federal statistics: combining data sources while protecting privacy*. Washington, DC: National Academies Press.

Prober, C. G., & Khan, S. (2013). Medical education reimagined: A call to action. *Academic Medicine, 10*(88), 1407–1410.

Reinsel, D, Gantz, J., & Rydning, J. (2017). *Data age 2025: The evolution of data to life-critical. Don't focus on big data; Focus on the data that's big.* IDC White Paper. Retrieved 2017 July 28 from https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf

Report of the Commission on Evidence Based Policy Making (2017): *The promise of evidence-based policymaking*. Washington D.C. Retrieved 2018 January 4 from https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf

Rigobon, R. (2015). Discussion on applications and issues with using commercial data in research. BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics, November 19, 2015. Washington D.C. Slides retrieved from http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_170272.pdf

Samoilova, E. (2017). Program assessment report: Flipping classroom in online courses for working professionals: challenges and opportunities for student engagement *International Program in Survey and Data Science. Assessment Report #2.* Retrieved 2017 November 1 from https://survey-data-science.net/sites/default/files/ipsds_flipping_classroom_in_online_courses_for_working_professionals-_challenges_and_opportunities_for_student_engagement_0.pdf

3 Sociocultural perspectives

DATA SCIENCE EDUCATION AS CONTRIBUTION TO MEDIA ETHICS

Tobias Matzner
Paderborn University
Warburger Str. 100
33098 Paderborn
Germany
tobias.matzner@uni-paderborn.de

INTRODUCTION

Current developments of information technology have widespread impact on our societies. Whenever ways to cope with these changes are debated, education or fostering competencies are part of the picture—for example under the rubrics of "media literacy" or "data literacy." A firm understanding of data, their meaning, the possibilities of deriving information from them—i.e. understanding important foundations of data science—are held to be a central element of such an education.

This means that data science education from an early age is not just considered to be important because it gives students skills for the job market or the university. It is also held to be an important element of forming a person that can take rational and informed decisions in today's digital world. Understanding what data are and what can be done with them (and what cannot) is considered to be a relevant element of taking ethical decisions or forming political opinions.

In consequence, attempts to foster data science education, as an element in spreading "data literacy" or "media literacy" are paramount. The European Union recently organized a European Cyber Security Month, with the motto: "Stop Think Connect." The claim of this program is to "provide resources for citizens to protect themselves online, through education and sharing of good practices."[1] In Germany, the Federal Office for Information Security (BSI) made education a prime part of its activities, launching a new website on information security for citizens.[2] This is a decisive step away from the former orientation of the office, whose central activities were counseling political institutions and big enterprises. Another example is the recent open letter of the Privacy Commissioner in Canada, stating that:

> "a recent survey […] found 92 per cent of Canadians are concerned about the protection of their privacy and nearly half feel as though they've lost control over how organizations collect and use their personal information.
> These findings are alarming and can only be addressed if we help younger generations to develop skills that will allow them to navigate the ever complex digital environment safely and responsibly."[3]

This emphasis on individual citizens' skills is particularly illustrative since it comes from a supervisory authority with the right to sanction corporate and state actors. Furthermore, the Privacy Commissioner in Canada, in alliance with the Privacy Commissioner of Ontario, has been one of the most outspoken advocates of privacy by design.[4] This concern with regulation and the design of products, however, seems no longer to be a sufficient scope for privacy protection. The examples I have listed point at one of the central issues of data literacy as a solution to ethical and political problems: Where to locate the responsibility for sensible uses of data—at the individual, the state, or the corporate level? I cover this question under the rubric of responsibilization below. The other issue I want to focus on is the rationalist bias of many current outlooks on media ethics. First,

[1] https://cybersecuritymonth.eu/about-ecsm/whats-ecsm
[2] https://www.bsi-fuer-buerger.de
[3] https://www.priv.gc.ca/en/opc-news/news-and-announcements/2017/nr-c_171108/
[4] https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

however, I want to quickly motivate why such social and/or political outlooks are important even for data science classes, which mainly focus on technical and mathematical skills.

IMPLICATIONS OF DATA SCIENCE EDUCATION

Even if an education program focuses on technical skills like data analysis, it can convey implicit messages. The use of "implicit message" here is meant to gather a group of phenomena which have been analyzed as "hidden curriculum" (Kentli 2015) or various forms of implicit biases, without following a specific theory of education. In particular, helping students understand why the contents of a course are important for them, is part of most classes. Thus, when teaching data science skills, it is likely that teachers will at least implicitly suggest why data science skills are important for everyone, beyond the more specific issues of preparing for particular professional careers. To illustrate my point, I want to distinguish two kinds of reasons why data science skills could be important for students.

On one hand, one might argue that data science skills are necessary for making informed decisions regarding the individuals' own actions. Such a perspective would suggest that it is necessary to understand the consequences of using digital data when, e.g., using your phone, watching a movie online, navigating the city with digital maps, reading the news, shopping online, etc. Here, the students are acting as as *rational individuals*.

On the other hand, one might argue that data science skills are important to judge social, technical, and political developments and agendas, i.e., not just an individual's own actions but their take on society and its technological conditions. This perspective would suggest that it is necessary to understand the potential for using digital data when, e.g., judging and discussing political programs, supporting an NGO, thinking about what you read in the news, debating privacy and security at work, or pondering whether you should be bothered to care—or not. Here students are acting as as *informed citizens*.

Of course, this distinction is by no means mutually exclusive and is meant as an idealized way to illustrate different groups of aims. The issues I discuss, responsibilization and rationalist bias, might be summarized as posing the danger that classes focus only on rational individuals, forgetting the perspectives on informed citizens. Thus, I argue not so much for replacing one for the other, but by amending the first perspective with the second.

RESPONSIBILIZATION

Responsibilization is a term from political theory, which means "the process whereby subjects are rendered individually responsible for a task which previously would have been the duty of another—usually a state agency—or would not have been recognized as a responsibility at all" (O'Mailey 2009, p. 276). The term has its origins in debates on criminality and security (Garland 1997), but has been extended to other areas. For example, the privatization of healthcare, pension plans, etc. are a responsibilization of welfare. Together with colleagues, I have applied the concept of responsibilization to data protection (Matzner et al. 2016). Here, I extend this outlook to data science skills or data literacy more broadly. In this case, responsibilization suggests that individuals are responsible for the data they provide, use, or refuse to share. Doing so in a rational manner requires a profound knowledge of data, of the ways it can be analyzed, and of the information that can be gathered from it.

This focus on the individual is problematic in various areas where data is used. In many cases, even with profound knowledge, the implications of providing data are very hard to judge: digital data are easily stored and have unforeseeable future uses; algorithms from machine learning, which have become more common in data science, are very hard to scrutinize; and even if it were possible, they are often protected as trade secrets. In consequence, responsibilization confronts individuals with the burden of a Sisyphean task, which means that digital environments (i.e., all environments) can become menacing or unpleasant. This is especially true when teaching focuses on understanding data and their meaning rather than addressing non-cognitive dimensions of learning, like teaching strategies and practices. This can evoke the feeling of hopelessness: if all that reckoning is necessary whenever I use my phone, I might as well accept that attempting a rational handling of data is futile.

Furthermore, even if we managed to provide excellent data science skills to everyone, relegating the responsibility for data to individuals who provide them is still prone to being unjust. In particular, the burden to care for the privacy and security of one's data is unequally distributed along well-known lines of discrimination, like gender (Cheney-Lippold 2011), skin color, citizenship (Adey 2012; Leese 2014), legal status (asylum seeker, undocumented, …), or dependence on welfare (Gilliom 2001). Thus, caring for the possible consequences of providing data has different salience for different social groups, and different amounts of work, skill, and attention are necessary for each group. Thus, the formally equal requirement that everyone be responsible for their own handling of data creates very unequal outcomes.

Finally, individual responsibility for data contradicts one of the formal insights of data science: Data about one group of individuals can be generalized and used to predict information about others (Matzner 2014). Thus, the idea that everyone should care for their own data presupposes a possibility to partition data into individually-relevant portions—but such a partition may no longer be available. (This fact also creates huge problems for legal frameworks based on individually identifying information (Nagenborg 2017)).

RATIONALIST BIAS

The second issue pertains to the implicit—or often explicit—demands that students should be able to make rational decisions regarding their use of data or information technology more broadly. Gather knowledge, think carefully, know what you do, seem to be the central lesson of many data literacy classes—in short: be rational. This view is not just common in data literacy or media literacy programs. Many academic analyses of people's usage of information technology are explicitly framed in rational choice models or related perspectives like e.g. the "privacy calculus" (Keith et al. 2013). Also, many big enterprises argue that they offer a product and the users should judge—or rather calculate—if the data they "pay" is worth the service they get. Data literacy is seen as one way to perform this gauging in a rational manner.

However, digital technology is part of our daily lives, not just the rational part of it. The appeal to rationality precludes spontaneous, emotional media use—a huge factor in the success of social media. If media ethics just means acting more rationally, it might not do justice to the reality of peoples' lives. Classes that suggest as much might be considered very unworldly, especially for teens (Marwick and boyd 2014).

More importantly, the implied division of rational and emotional or affective decisions has been seriously challenged across the sciences and humanities. Results from biology and neuroscience, maybe most prominently Damasio's work (Damasio 1994), have inspired the "affective turn" in cultural studies (Clough and Halley 2007). The perspectives in this field differ widely and are strongly debated amongst each other (Leys 2011; Wetherell 2013) as well as regarding their relationship to earlier views like psychoanalysis (Angerer 2014). Yet, they all share the view that a strong distinction in rational and emotional or affective cognition and behavior is not tenable. This insight has also motivated the emergence of new strands of research in computer science, like affective computing (Picard 2003; Picard 1997), which makes affect an integral aspect of human-computer interaction. Of course, designing products in order to elicit particular affective reactions—in particular longer periods of use—is an important aim in creating digital products, even attempting to create "habit forming" (Eyal 2014) or "addictive" (Dow Schüll and Zaloom 2011; Dow Schüll 2005) products. Recently, a group of former top Silicon Valley executives has founded the NGO "Humane Tech" to raise awareness to the psychological and emotional "manipulation" of our brains by IT companies.[5] Even if they remain in a strongly technologically deterministic setting and thus maybe attribute too much power to technology—like almost direct access to our "brains"—it is noteworthy that even Silicon Valley entrepreneurs start to be worried about the affective influences technologies have on our lives.

In such a context, demanding that students decide rationally by understanding the features and potential uses of data, or by rationally choosing among different offers, obscures all these affective influences. A truly rational approach to digital technology thus needs to include a self-understanding that does justice to the constitutive role affect and emotions have for our

[5] www.humanetech.org

subjectivity. Of course, data science classes alone cannot achieve this; this is something that probably needs to be taken up across all subjects in school. But it is important for data science classes not to suggest that students should learn data science as the basis for rational self-control—which they will consequently find almost impossible to achieve in their life world, which is permeated by affective triggers.

CONCLUSION

Both issues, responsibilization and the rationalist bias, underline the importance of data science education. However, they also show that such education needs a careful framing regarding why students need such an education and which reflections and actions are required of them. Such a framing is often transported in the subtext, influencing what students think is required of them if they have to learn data science as part of becoming a reflective and ethical media user and citizen. Thus, I want to suggest that data science education needs to be embedded in teaching social and economic contexts, rather than focusing on the individual students and their individual skills and actions. Technical knowledge and insight are essential for understanding most of these aspects. However, they need to be integrated with a broader view, at best in an interdisciplinary setting. Some questions that could be addressed along these lines are: Who are powerful actors in the field of data? Who has huge resources of data? What can they do with data, which I/we cannot? Why is a question solved with data science in the first place? Which would be the alternatives? Which social practices give data its specific meaning and relevance?

In such a context, data science education can play an essential part for education more broadly, aiming not just at rational individuals but informed citizens.

REFERENCES

Adey, P. (2012). Borders, identification and surveillance. In: Lyon, D., Ball, K. & Haggerty, K. D. (eds.) *Routledge Handbook of Surveillance Studies*. London: Routledge, pp. 193–201.

Angerer, M.-L. (2014). *Desire after affect*. Rowman & Littlefield International.

Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society, 28*, 164–181.

Clough, P. T. and Halley, J. O., eds. (2007). *The affective turn: Theorizing the social*. Durham: Duke University Press.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.

Dow Schüll, N. (2005). Digital gambling: The coincidence of desire and design. *The Annals of the American Academy of Political and Social Science, 597*, 65–81.

Dow Schüll, N. and Zaloom, C. (2011). The shortsighted brain: Neuroeconomics and the governance of choice in time. *Social Studies of Science, 41*, 515–538.

Eyal, N. (2014). *Hooked: How to build habit-forming products*. New York, New York: Portfolio/Penguin.

Garland, D. (1997). 'Governmentality' and the problem of crime: Foucault, criminology, sociology. *Theoretical Criminology, 1*, 173–214.

Gilliom, J. (2001). *Overseers of the poor: Surveillance, resistance, and the limits of privacy*. The Chicago series in law and society. Chicago: University of Chicago Press.

Keith, M. J., Thompson, S., Hale, J., Lowry, P., & Greer, C. (2013). Information disclosure on mobile devices: Re-examining privacy calculus with actual user behavior. *International Journal of Human-Computer Studies, 71*, 1163–1173.

Kentli, F.D. (2015). Comparison of hidden curriculum theories. *European Journal of Educational Studies, 1*.

Leese, M. (2014). The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue, 45*, 494–511.

Leys, R. (2011). The turn to affect: A critique. *Critical Inquiry, 37*, 434–472.

Marwick, A.E. and boyd, d. (2014). Networked privacy: How teenagers negotiate context in social media. *New Media & Society, 16*, 1051–1067.

Matzner, T., Masur, P., Ochs, C. and von Pape, T. (2016). Do-it-yourself data protection: Empowerment or burden? In Gutwirth, S., Leenes, R. & De Hert, P. (eds.), *Data Protection on the Move* (pp. 277–305). Dordrecht: Springer.

Matzner, T. (2014). Why privacy is not enough privacy in the context of 'ubiquitous computing' and 'big data.' *Journal of Information, Communication and Ethics in Society, 12*, 93–106.

Nagenborg, M. (2017). Informationelle Selbstbestimmung und die Bestimmung des Selbst. In Friedewald, M., Lamla, J. & Roßnagel, A. (eds.), *Informationelle Selbstbestimmung im digitalen Wandel* (pp. 65–72). Wiesbaden: Springer Fachmedien.

O'Mailey, P. (2009). Responsibilization. In Wakefield, A., & Fleming (eds.), *The SAGE Dictionary of Policing*. London: Sage.

Picard, R.W. (1997). *Affective Computing*. Cambridge, Mass: MIT Press.

Picard, R.W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies, 59***,** 55–64.

Wetherell, M. (2013). Affect and discourse—what's the problem? From affect as excess to affective/discursive practice. *Subjectivity, 6*, 349–368.

ALGORITHMIC LITERACY

Katharina A. Zweig, Tobias D. Krafft,
Sujay Muramalla, Julien Siebert
Gottlieb-Daimler-Str. 48, 67663 Kaiserslautern
{zweig, krafft, muramalla, siebert}@cs.uni-kl.de

In different contexts, decisions are more and more delegated to algorithms. Humans (domain experts, policy-makers, etc,) need not only to understand the computer-generated results but also to make informed decisions about them. An end-to-end understanding of algorithmic decision-making (ADM) systems requires a set of various skills and knowledge from several domains. We refer to this set of competencies as *algorithmic literacy* and propose a systematic review of this new literacy. Recognizing "algorithmic literacy" is just the first step of developing a methodology for assessing the future impact of ADM systems before they are employed—similar to, for example, environmental impact studies before new building or plants are built.

Algorithms have taken over many decisions that were formerly made by humans: Where a bookseller would help us to find a new book to read or an employee of a video store would recommend us the next movie, today we have Amazon and Netflix do this through so-called recommendation systems. A recommendation system tries to learn from past data by deducing rules that predict which product a given customer might like. This is done by finding correlations in past data, such as: "people like you liked…" or "this product was often liked together with …". Most often, such systems use machine learning algorithms to find and store the rules. According to Jannach et al. (2012), recommendation systems on the one hand solve an *information retrieval* problem, namely to equip you with the most relevant items. On the other hand, the authors emphasize, there is a machine learning perspective, as the systems try to learn a model and to *predict* the (rating) behavior of the customer. Once the prediction has been made, those items with the highest predicted rating will be recommended. Last but not least, Jannach et al. stress that the system is one type of an *algorithmic decision-making (ADM) system* which supports the decision of humans. While there are ADM systems without a machine learning component, e.g., those who get their decision rules from a set of experts, ADM with a machine learning component are the most troublesome (Zweig, Fischer & Lischka, 2018), so, in the following, we will focus on them. Additionally, ADM systems can make decisions on all kinds of objects and subjects. For this paper, we will solely focus on ADM systems with a machine learning component that classifies or predicts the behavior of humans, because they risk potential damage to individuals, and thus require a very careful interpretation of the machines' results (Mittelstadt, et al. 2016).

All of these ADM systems suffer from the same problems that influence the result's quality and interpretability. The first class of problems lies in the **collection and selection of the data** from which the machine learning algorithm learns the rules, also known as *training data*:

1) The training data can be erroneous. Especially, when data comes from unstructured sources or is composed from multiple databases, it is likely that information belonging to two different entities are merged or that information about the same entity is regarded as being from different entities (the so-called entity recognition problem). Other problems are outdated information or missing information. For various other kinds of "bad data" see McCallum (2012).
2) Often, the system learns from the decisions of human experts, for example, when the system tries to find rules for what makes a job applicant likely to be successfully hired. If these previous decisions were discriminating, the ADM results will also be discriminating.
3) Very often, an important dimension of information cannot be directly assessed. For example, if a certain behavior partly depends on (real-world) friends' behavior, it is necessary to determine who is a friend of whom. However, that information is very difficult to get, so data scientists tend to use digitally accessible information such as online friendship, or following-relationships on a social media platform, as a proxy of real-world relationships.

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

We can thus describe the first competency that a user needs to be algorithmically literate:

Competency 1: Evaluation of quality of training data

A person who needs to interpret the result of an ADM system with a machine learning component needs to know of possible problems within the training data. He or she also needs to understand the sensitivity of the machine learning algorithm to potential errors contained in the data.

This first competency is not specific to ADM but is also required from anyone that needs to interpret statistical results. The reader interested in statistics methodology may refer to Reinhart (2015) and to McCallum (2012) for a more ADM-oriented point of view on these problems.

The next two classes of problems that can arise in the interpretation of an ADM result are rather technical:

1) **Problems in the implementation phase:** Is the machine learning algorithm implemented correctly, i.e., are the rules deduced from the data in the anticipated way?
2) **Problems while training the ADM system:** Is the right machine learning algorithm used in the ADM system?

Both questions require profound technical skills and can and should be answered by computer scientists, software engineers, or *data scientists*. Implementation problems are actually well-covered in any traditional software engineering textbook. The possibility of finding technical implementation mistakes depends essentially on three aspects: Do a lot people use the algorithm? Was the behavior of the algorithm well specified? Is the source code publicly available? Most programming bugs are detected fast when these three aspects are met.

The second class of problem is primarily dealt with by the data scientist—with the help of 'domain experts'—and competencies from both are needed. The data scientist is the person that combines method and data and is often a trained physicist, mathematician, computer scientist, or statistician (Dhar 2013). However, there is no classical curriculum for this new profession; often data scientists come with some technical background knowledge about various methods, but lack the experience to tackle problems from the domain in which the ADM system is applied. For example, if a data scientist is asked to design a *recidivism risk assessment tool*, he or she would need the insight of various domain experts (criminologists, sociologists, psychologists, etc.) to understand which algorithmic method is most likely to produce the most insight into the problem. From this we deduce the second necessary competency:

Competency 2: Interdisciplinary communication

At some point in the development of an ADM system, the developers (called 'data scientists') and the experts in the field (called the 'domain experts') need to be able to discuss with each other. Important questions are: what kind of data should go into the system and what should not, is the behavior to be predicted a multi-causal behavior (which would exclude some machine learning methods), and what are the main known causes for the behavior? The data scientists need to ask questions like how high the error rate is on the data or whether there is missing information. The domain experts need to be able to give a quantifiable measure on when the behavior to be predicted occurs and when not. E.g., if the ADM system is asked to predict recidivism, it needs data that contains information about whether somebody recidivated, which requires a clear definition of that behavior. The data scientists then decide which data and which method to use. In a final step, data scientists and domain expert need to agree on how to rate the quality of an ADM system while it is trained. As we showed earlier, this quality measure depends—among other things—on the social process the ADM system is embedded in (Krafft 2017).

As a last step of the ADM design, the output of the system needs to be decided upon. In many cases, the ADM system will assign a number to each individual, a number that represents the likelihood of the person to show the behavior being predicted. Then the persons can be ordered

according to this probability. Very often, simpler representations of the initial result are computed, for example, by assigning people to a smaller set of classes based on the initial ordering (see Figure 1).

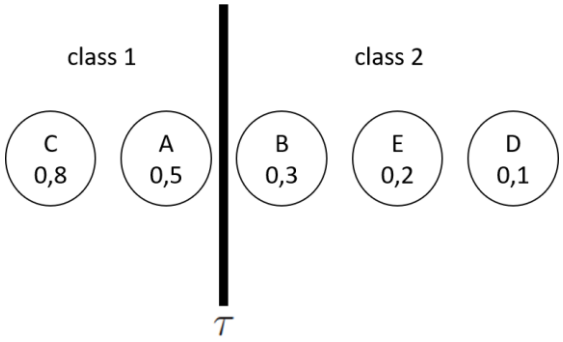


Figure 1: The ADM has assigned a probability of "membership to class 1" to the elements (A-E). If sorting is now done and a threshold $(\tau = 0{,}4)$ *is selected, the visualized division into classes 1 and 2 results. Any element with a probability of "membership to class 1" greater than* $\tau = 0{,}4$ *is then classified into class 1.*

The American Civil Liberties Union (2011)—which advocates the use of risk assessment systems in every step of the American legal process—proposes a classification of criminals in three risk classes: high risk for recidivism, medium risk, and low risk for recidivism. But various factors might bias the perception of users, such as using colors or specific symbols (Easterby, 1970; Carrillio et al. 2014) or by using conditional probabilities that are very often misunderstood (Hoffrage et al. 2002). From this, we deduce competency number 3:

Competency 3: Knowledge about human biases in interpreting statistics and in decision-making

A person who needs to interpret the result of an ADM system with a machine learning component needs to know how he or she is biased by a certain representation of a result.

The next class of problems comes from the interaction between the ADM and the social context it is embedded in. Because it contains a learning component, phenomena that are difficult to foresee may occur due to feedback loops between society (or the group of impacted people) and the ADM process. In her book "Weapons of Math Destruction" (2017), Kathy O'Neil uses the example of predictive policing. Predictive policing is based upon the "near-repeat" theory that states that some types of crimes are likely to happen in the same area for a given period of time. The outcome of a predictive policing ADM can lead to more patrols in some areas. This in turn automatically leads to more arrests in the area as more petty crime is discovered. As a result, the system "learns" that many criminals live here, which could further increase the number of patrols in these areas, etc. This example illustrates that, because of a self-reinforcing feedback loop, a seemingly objective measure presents a risk of changing how crime is pursued, in an uneven way, and can therefore give the false impression that a subgroup of the population is much more criminal than the rest of the population. This example illustrates the need of the following competency:

Competency 4: Knowledge about potential impact and possible feedback loops.

A person who needs to interpret the result of an ADM systems with a machine learning component has to understand that existing self-reinforcing feedback loops may render ADM results obsolete. He or she needs to know whether such feedback loops exist and whether their impact has been mitigated or not. This last competency implies skills and methods originating in decision-making (Brest & Krieger 2010) and complex system theory (Dekker 2011).

In summary, algorithmic literacy is more than just the juxtaposition of skills—it's the coordination of different competencies towards the goal of interpreting and assessing ADM results. Furthermore, the increasing number of incoming recommendations and laws about ADM will require new functions and jobs, for which more detailed scientific research with this new literacy is necessary. We highlight the potential and the need of new kinds of training and formation. This systematic approach, with different competencies referencing various problems and research fields, is a first step in the direction of a methodology for assessing the impact of ADM.

References

American Civil Liberties Union (2011). *Smart reform is possible: States reducing incarceration rates and costs while protecting communities*. Technical report. Retrieved Feb. 26, 2018 from https://www.aclu.org/smart-reform-possible-states-reducing-incarceration-rates-and-costs-while-protecting-communities

Brest, P., & Krieger, L. H. (2010). *Problem solving, decision making, and professional judgment: A guide for lawyers and policymakers*. Oxford University Press.

Carrillo, E., Fiszman, S., Lähteenmäki, L., & Varela, P. (2014). Consumers' perception of symbols and health claims as health-related label messages. A cross-cultural study. *Food Research International*, *62*, 653–661.

Dekker, S. (2011). *Drift into failure: From hunting broken components to understanding complex systems*. Boca Raton: CRC Press.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73.

Easterby, R. S. (1970). The perception of symbols for machine displays. *Ergonomics*, *13*(1), 149–158.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, *84*(3), 343–352.

Jannach, D., Zanker, M., Ge, M., & Gröning, M. (2012). Recommender systems in computer science and information systems–a landscape of research. *International Conference on Electronic Commerce and Web Technologies*, 76–87. Springer, Berlin, Heidelberg.

Krafft, T. D. (2017). *Qualitätsmaße binärer Klassifikatoren im Bereich kriminalprognostischer Instrumente der vierten Generation*. Master thesis, Fachbereich Informatik, TU Kaiserslautern. arXiv preprint arXiv:1804.01557.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. London: Penguin.

McCallum, Q. E. (2012). *Bad data handbook*. O'Reilly Media, Inc.

Reinhart A. (2015). *Statistics done wrong: The woefully complete guide*, first edition. San Francisco, CA: No Starch Press.

Zweig, K.A., Fischer, S., & Lischka, K. (2018). *Wo Maschinen irren können – Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung*. Bertelsmann Stiftung, https://doi.org/10.11586/2018006

4 Data science in institutions, media, and companies

PERSPECTIVES FOR DATA SCIENCE EDUCATION AT THE SCHOOL LEVEL: HOW TELEKOM GERMANY DRIVES AWARENESS FOR BIG DATA AND DATA DRIVEN OPPORTUNITIES

Christoph Denk
Head of CRM Data Management
Landgrabenweg 151
53227 Bonn
Telekom Germany
Christoph.Denk@telekom.de

This brief paper reports some reflections regarding the teaching of data science at the high-school level. The paper highlights the "big data and data analytics awareness program" of Telekom Germany.

INTRODUCTION

Deutsche Telekom offers services for

- 165 million mobile customers
- 29 million fixed-network customers
- 19 million broadband customers
- Approx. 6.9 million TV customers[1]

Each customer has the expectation that Deutsche Telekom offers a convergent and a convenient service experience via all customer touch points to meet individual expectations. Data is the resource and the key asset for meeting customer expectations in a digitized world. Telecommunication industry trends such as the "internet of things" require even more data to offer a sophisticated customer experience.

Data analytics is a cornerstone in the strategy of Telekom Germany and it is important that our employees are aware of the benefits and opportunities associated with data, the use of data, and the data business. Statistics and programming are not professions for everyone, but everyone needs to understand how the industry works and operates. Our digital life is based on big data. This concerns our whole society. The understanding of big data should be a basic skill for everyone.

The "big data and data analytics awareness program" of Telekom Germany is designed to meet the needs of our company. Therefore, only certain parts of the program are adoptable at the high-school level.

GET EXCITED, GET INSPIRED, GET STARTED

The Telekom Germany data analytics awareness program is based on various formats and starts with a web-based training that is accessible to every employee.

The web-based training starts with the GET EXCITED chapter to explain the concept behind big data and data analytics. It doesn't matter if you use individual entertainment services, convenient same-day delivery services, new payment services, or streaming technology. The digital revolution and big data enable innovative business models that were impossible in an analog world. New players enter the market and attack the established economy with personalized and convenient offers for everyone (e.g., Uber offers transportation without owning their own cars, Amazon offers books without printing.)

[1] Source: DT 2016 annual report/TMUS annual report to shareholders 2016

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

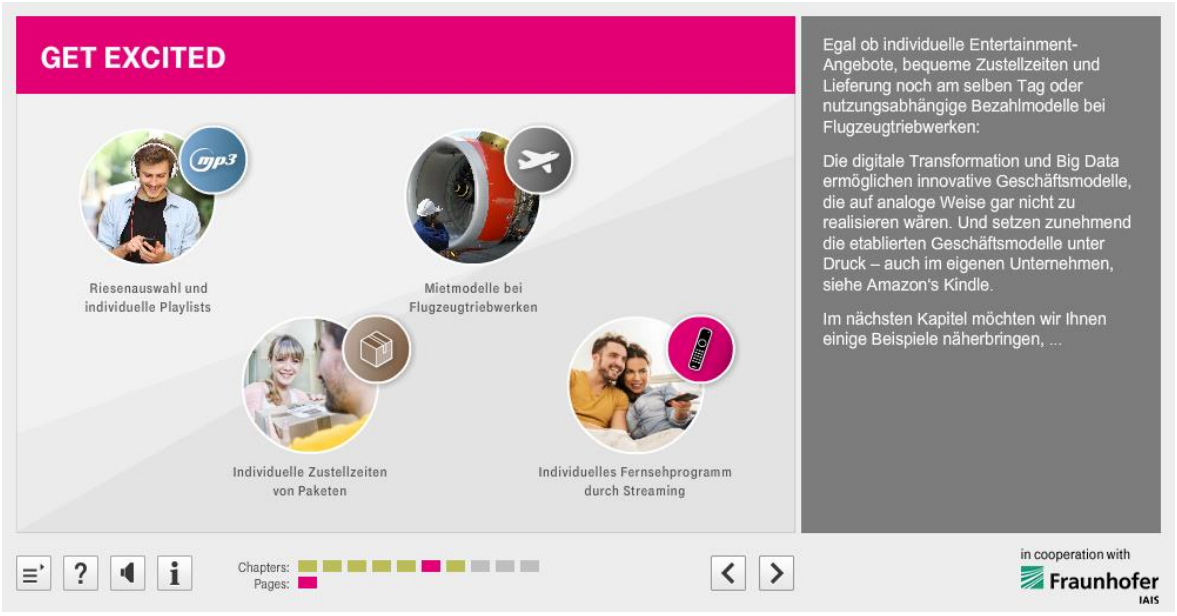


Figure 1: Web based training - GET EXCITED

The chapter GET INSPIRED explains four trends: convergence, smart systems, digital lifestyle, and open data, which lead to big data. Digital convergence, the expected explosive distribution and use of intelligent interfaces, and the digitization trend of our whole society lead to a long-term and growing mass of data. The mass of data in turn leads to new services, generating new data, which lead to business models that rely on generated data (e.g., the use of search engine phrases for user-centered advertisement). The merger of these trends generates pressure on established businesses. Today's product and service development requires enthusiasm for innovation and a constant change of existing working modes. Agile co-working between different disciplines within our organization is a key to adopting new trends quickly.

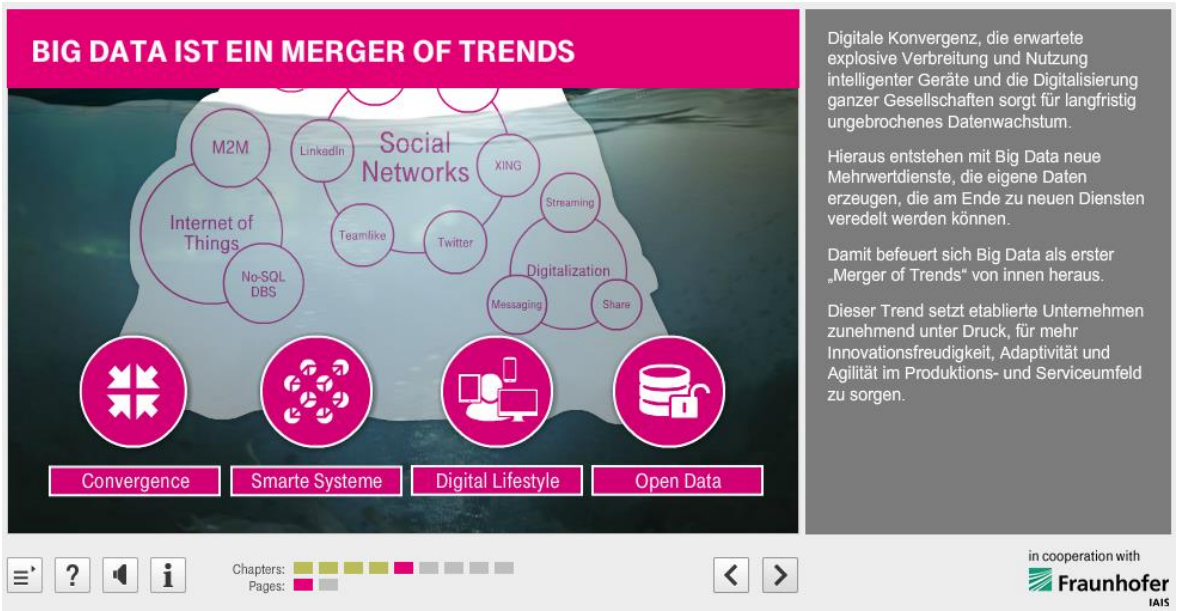


Figure 2: Web based training - GET INSPIRED

The GET STARTED chapter of our web-based training introduces the new job profiles in a data scientist team:

a) business expert
b) big data analyst
c) data engineer

Telekom Germany needs business experts who are aware of the big data potential for their business. They need to identify business opportunities and they have to ask the right questions to drive their business. The big data analysts map the business requirements for existing and new data, and must be able to visualize the business problem. The data engineers build the data pool from all data sources to fulfill the needs of the big data analysts. Telekom Germany offers learning opportunities for different skill levels for each job profile.

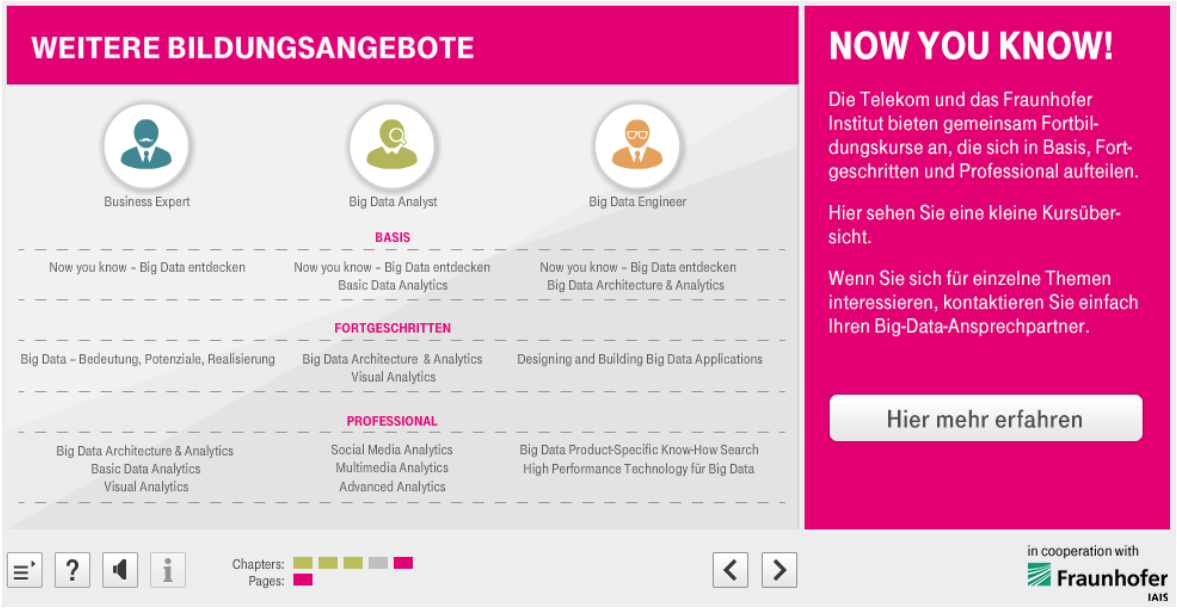


Figure 3: Web based training - GET STARTED

Besides the web-based training, Telekom Germany also offers workshops which are integrated into team meetings or business conferences. These workshops start with an introduction of the term "big data" from industries other than telecommunications, and focus on the potential of big data. The health care industry is a good example for reflecting on data privacy constraints (e.g., using smart watches and fitness apps), perceptions of what individuals want to share, and the industry perspective on digitization. The example and the discussion with the participants lead to three questions which limit the use of big data.

1. What is technically possible—including whether the data stream is machine readable and available?
2. What is legally allowed?
3. What is acceptable for the user?

The discussion of the three questions is the starting point for the participants to develop their own big data business cases based on data from their own department.

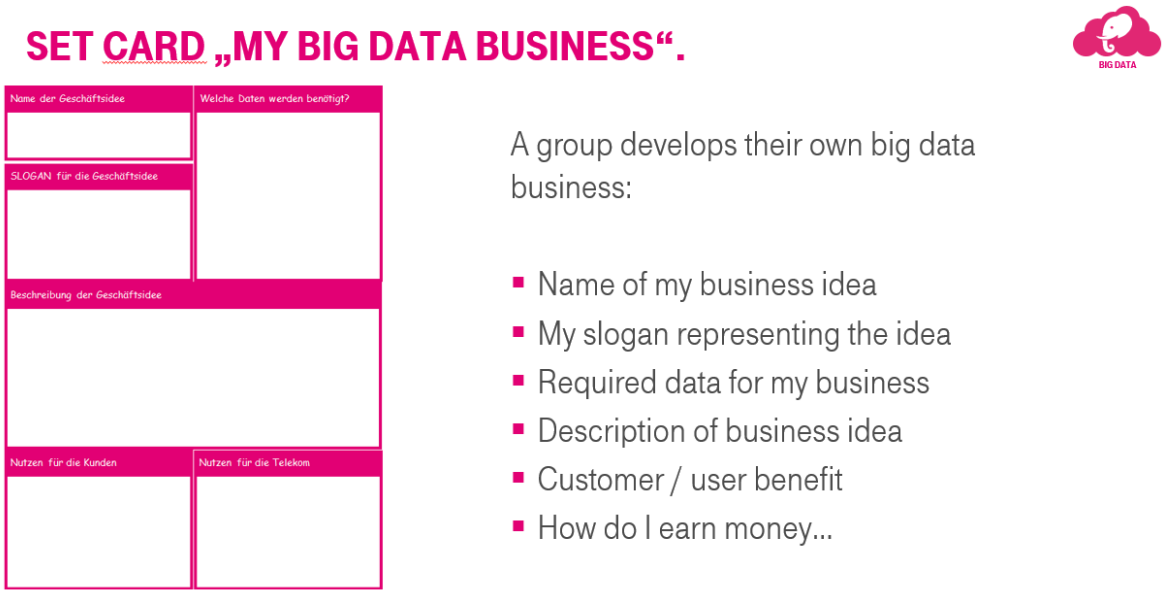


Figure 4: Set card - Big Data business idea

The introduction of the term big data is important to awaken the interest in big data, the business models, and the technology. An introduction to big data at school level has to ensure that every student is familiar with the basic concept of big data:

- big data concerns everyone
- big data is a business
- data is an asset
- the limiting factors of big data
  a) What is technically possible?
  b) What is legally allowed?
  c) What is acceptable for the user?

FROM EDUCATION TO NANODEGREE

If the employees of Telekom Germany want to learn more, they can make use of our education program. Telekom Germany developed it together with the Fraunhofer Institute. The trainings are focused on the roles in a data scientist team. The business expert with the subject matter expertise defines the required business improvements. The big data analyst needs skills in data mining, machine learning, and text mining. The big data engineer needs IT architecture and deployment skills for our Hadoop platforms.

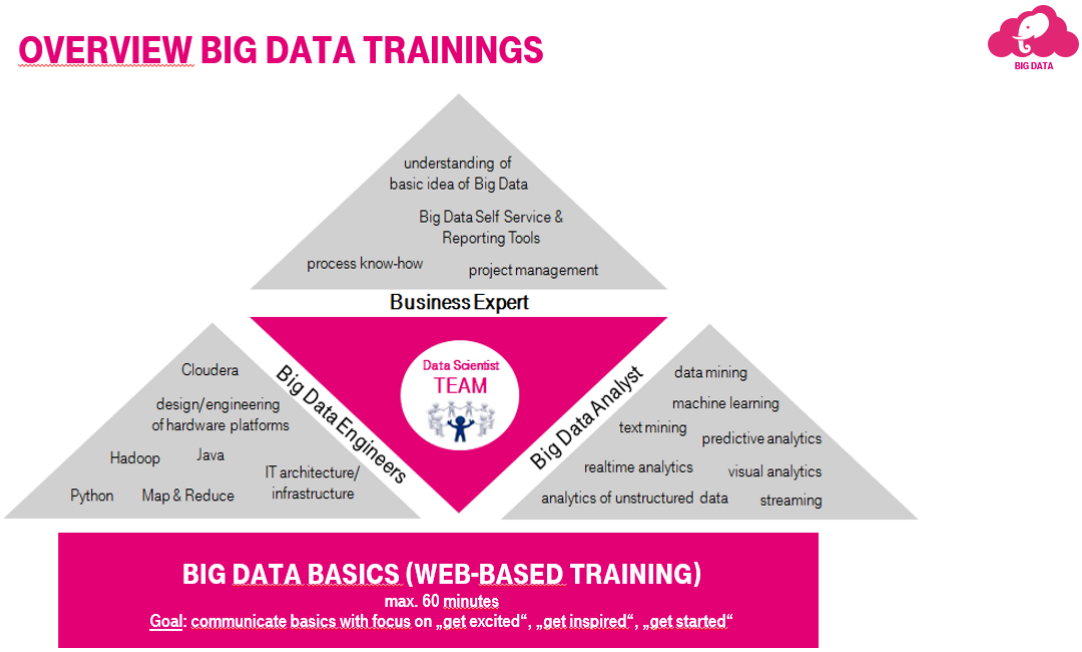


Figure 5: Big Data trainings for different target groups

Programming experts asked for a training program that is more focused on programming details and new tools. Telekom Germany offers the opportunity to obtain a nanodegree in data science in co-operation with Udacity. Our data analyst community can sign up for a one-year e-learning course where the employee receives a data analyst nanodegree, a generally-accepted certificate to prove your qualification. Parts of this training might fit into advanced computer science courses.

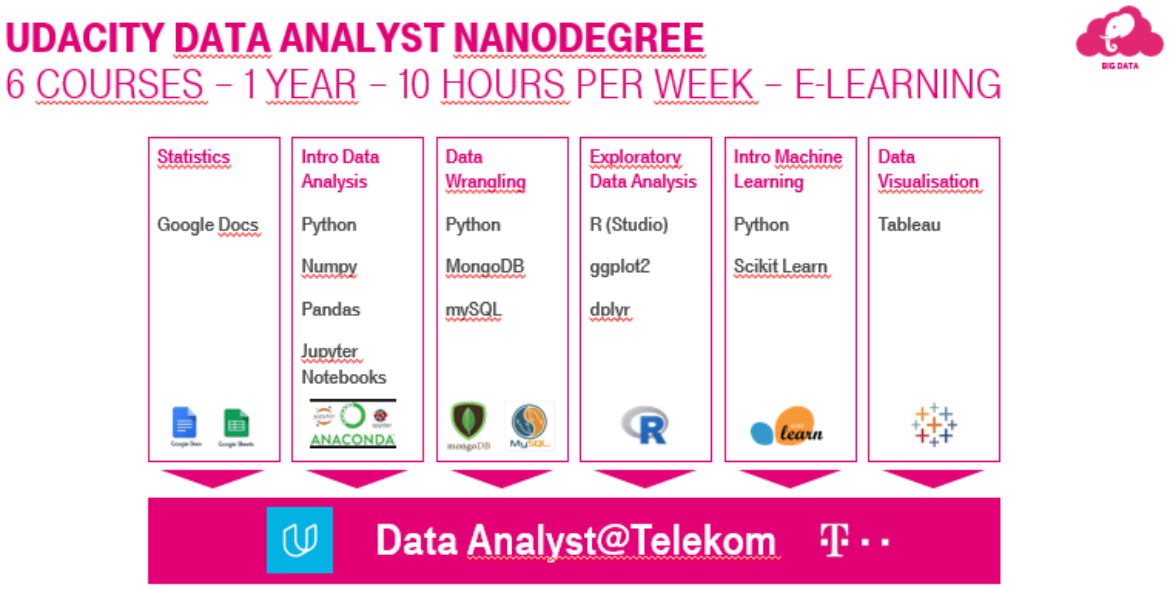


Figure 6: Udacity training program

A lot of discipline is required to finish the one-year program. At least 10 hours per week are necessary in order to finish the program in a year, and it's not easy. Everyone can sign up for the course. The only prerequisite is a basic math and statistics background and the willingness to learn programming including the willingness to invest one's own personal time. The only infrastructure requirement is a web browser to access the course, and the student is able to do the course when he has time and is able to define his own learning speed. Each learning topic ends with a short questionnaire to verify that everything has been well understood. Each course ends

with a practice test, where the student receives a data set and a business question to solve. The individual solution is assessed by a Udacity coach and each successful practice test is honored with a certificate. The sum of the certificates constitutes the nanodegree which proves the capabilities of the student.

Understanding data science and big data also requires the teaching of at least one programming language for data wrangling and data analytics. Python is a good starting point to learn a data science programming language at the school level. It's easy to learn Python. Many online tutorials are available for free. It's also very easy to set up a Python environment on a personal computer.

COMMUNITY BUILDING

The presentation of success stories is important for a specialized community to show their results to a broader audience. There are art exhibitions at school and also the school band and the chorus are on stage. Even the school sports teams have the ability to compete with other teams.

A data scientist community also needs a regular personal exchange and they also need events to celebrate success stories and to receive acknowledgment for excellent performances. Twice a year, Telekom Germany celebrates these success stories in an exhibition.

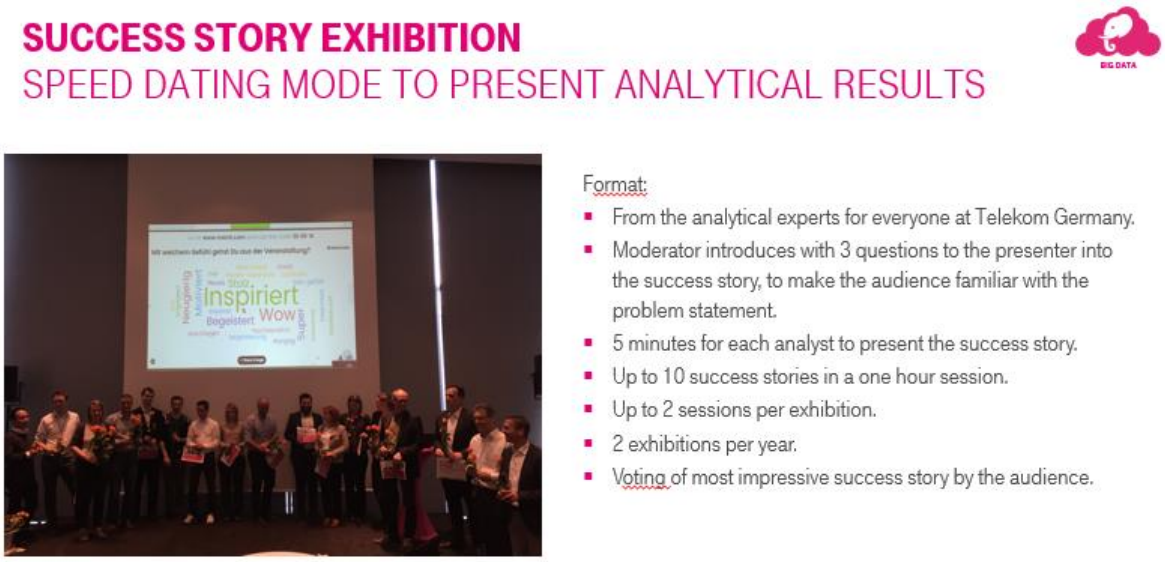


Figure 7: Communicate success - speed dating format

CONCLUSION

The Telekom Germany success story to drive awareness for big data started with the introduction of the term in basic workshops and an online seminar. Our second step was the development of a training program with our partners. The third step is the execution of the training program and the daily work with new technology. And last but not least, we market our success stories with our analytical community to a broader audience.

DATA SCIENCE IN THE NEWS BUSINESS

Andreas Loos
ZEIT online, Askanischer Platz 1, 10963 Berlin
andreas.loos@zeit.de

ZEIT online is one of the largest online newspapers in Germany. Uncommonly for a news site, ZEIT online has a team of mathematicians and data scientists to support investigative and data journalists by cleaning, classifying, searching, and processing/analyzing data from heterogeneous sources. We will present some examples from this work that could be reproduced or redone in a similar way at high school.

CORRELATIONS BETWEEN ELECTORAL BEHAVIOR AND SOCIAL STRUCTURE

After the last Federal election in Germany, the electoral management body (Bundeswahlleiter) provided a small dataset containing the election results as well as information about the social structure of the election districts (Bundeswahlleiter 2017); the structural data originally comes from the German Federal Statistical Office (Destatis). For each electoral district, users can download the election results for each party as well as data on about 20 different statistical variables. These are, among others, the number of vehicle registrations, the number of privately shared homes, and the number of unemployed people.

After the 2017 Federal election, we computed at ZEIT online correlations for each pair of party result and structural variable and tested these correlations for significance. For this, we wrote an R-script that automatically selected and visualized the most promising cases (one criterion was a high coefficient of determination) to examine these cases later in detail by hand. With that technique, we found — among others — a clear correlation between the number of car registrations and the election results for conservative parties (CDU/CSU) (Blickle et al. 2017).

This project can easily be reproduced in school; the raw data is available as a csv table on bundeswahlleiter.de. The results can be used not only to explain what a correlation is, but also to show what it is not: for some variables, one can show that what appears at first glance to be a correlation is in fact an artifact. An example, based on differences in the electoral behavior between the eastern and western part of Germany, is the relation between the number of persons who receive social welfare benefits and the number of votes for a right wing party (AfD). Votes for the AfD and a lower number of welfare benefit receivers are both much more common in the eastern part of Germany. Inside both Eastern and Western regions, however, there is only a weak correlation (if at all) between the number of welfare benefit receivers and voters for AfD.

DETECTION OF TENDENCIES IN SURVEYS OF POLITICAL OPINIONS

Another project concerned tendencies in surveys of political opinions. This work extended previous work of Gregor Aisch (2013a) and was based on data from wahlrecht.de, which is a website that is logging the results of surveys of political opinions since 1994. Their dataset contains data from the seven largest survey institutes (Emnid, Infratestdimap, Forsa, GMS, Allensbach, INSA, and FGW). The idea of Gregor Aisch was to compare the mean results of each individual institute quarterly with the total mean or median of the quarterly means of the institutes to see if individual institutes under- or overestimate the polls for one or another party.

We enhanced Gregor Aisch's R-scripts (2013b) in many aspects and added fresh data from wahlrecht.de to the project; in total, by the end of September 2017, we had about 35,500 data points, data from political surveys of the last twenty years. One of our results was that the survey institute Insa systematically overestimated the AfD compared with the other institutes (sometimes by four percentage points) but at the same time proved to be more correct than other institutes when compared with the final election result (Loos 2017).

"DEUTSCHLAND SPRICHT" ("GERMANY TALKS")

A very successful project was "Deutschland spricht" ("Germany talks"), where readers of ZEIT online from all over Germany were invited to discuss, pairwise and face-to-face, about politics on the afternoon of June 18, 2017. The idea was to connect people with diverging opinions.

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

Therefore, the participants had to answer five binary questions in advance: "Does the West treat Russia fairly?" "Does Germany host too many refugees?", "Should Germany return to the D-Mark?", "Should homosexual partners be allowed to marry?", "Was it good to exit from nuclear energy?" The number of differing answers was taken as a measure for the divergence in opinion (Bangel, Faigle, & Loos 2017).

We fixed the maximum distance between participants at 20 km. We then built an undirected graph adding an edge if and only if the readers were living not more than that distance apart from each other (by estimating the distance from the postal code); we weighted each edge with the opinion-distance. Using a standard algorithm for the computation of maximum matchings implemented in the python-library *networkx*, we could produce a nearly perfect matching, and thus connected almost all participants to conversation partners with different political views. In the end, we set up about 2700 pairs.

The core idea of this project could be easily redone at school level for a meet-and-discussion project in the school; the implementation in python is straightforward.

CONCLUSION

The application of mathematics and computer science in journalism is a young and emerging field of work. While many of the projects in data journalism are based on confidential data and background knowledge, some of the projects are based on data that is publicly available; these projects can in principle be reproduced by everyone. Sometimes, this can already be done with mathematics and programming skills at high-school level, for instance in school projects for highly interested pupils. In the presented projects, pupils could not only learn about statistics and programming tools, but also how to analyze data and how to read results critically.

REFERENCES

Aisch, G. (2013a September 22). Wie tendenziös sind Wahlumfragen? *Zeit Online*. http://www.zeit.de/politik/deutschland/2013-09/wahlumfragen-parteilichkeit-bundestagswahl

Aisch, G. (2013b). *R-poll-bias: A set of R scripts to visualize and analyze bias in the polls*. Retrieved from https://github.com/gka/R-poll-bias

Bangel, C., Faigle, P., & Loos. A. (2017 June 18). Streiten Sie schön! *Zeit Online*. http://www.zeit.de/gesellschaft/2017-06/deutschland-spricht-teilnehmer-methode-ergebnisse

Blickle, P., Loos, A., Mohr, F., Speckmeier, J., Stahnke, J., Venohr, S. & Vollinger, V. (2017 Sept 24). Merkel-Enttäuschte und Nichtwähler machen die AfD stark. *Zeit Online*. http://www.zeit.de/politik/deutschland/2017-09/wahlverhalten-bundestagswahl-wahlbeteiligung-waehlerwanderung

Bundeswahlleiter (2017). *Bundestagswahl 2017 – Strukturdaten für die Wahlkreise*. https://www.bundeswahlleiter.de/bundestagswahlen/2017/strukturdaten.html

Loos, Andreas. (2017 Sept 26). Wie tendenziös waren die Umfragen? *Zeit Online*. http://www.zeit.de/politik/2017-09/meinungsforschung-bundestagswahl-umfragen-treffsicherheit

DATA LABS: NEW KNOWLEDGE FOR SCHOOLS?

Katharina Schüller
STAT-UP, Leopoldstraße 48, 80802 München
katharina.schueller@stat-up.com

It is widely accepted that digitization means far more than a technological change. We assume that statistical literacy serves as a success factor for the digital transformation process. But in Germany, statistics at school (and at universities) is still taught as if it were just a special case of mathematics. Students learn that statistical problems have a solution that is either true or false. They do not learn how to deal with probabilities, randomness and "dirty data." We therefore propose a novel way of teaching statistical literacy in Data Labs that are characterized by working on real-life problems with agile methods, professional support, and ongoing exchange and feedback.

INTRODUCTION: STATISTICAL LITERACY IN A DIGITIZED WORLD

In the age of digitization data seem to be ubiquitous: We produce them with our mobile phones, with every movement on the internet, via sensors or camera shots. In addition to the large "data leeches" like Google, Facebook & Co., cities also wish to have more access to our data. Platforms for reporting local problems like fixmystreet.com are supposed to improve the citizens' service, citizens' digital participation is intended to enable urban planning that meets their needs, while crowdsourcing—for example in the form of hackathons—may exploit the potential of Open Data (Schüller & Förster 2017).

We observe that data is both overvalued and undervalued. They are overvalued because the internet is full of outstanding examples in which data and data science can solve seemingly any common and not-that-common problem. The whole world gets excited about "the power of data." They are undervalued because data-generating systems provide endless amounts of bits and bytes that are neither linked nor organized and therefore cannot be turned into power. We still give away the bulk of our data without hesitation, in exchange for route planners, timetables, or the possibility to post pictures of our daily breakfast.

What happens with these data, and which risks and uncertainties are associated with data-based decisions is, however, hidden to most of us. Though science fiction writer H.G. Wells stated[1] about 100 years ago: "If we want responsible citizens in a modern technological society, we have to teach them three things: Reading, writing, and statistical thinking, that is the reasonable use of risks and uncertainties." Today, statistical thinking and data literacy are more important than ever.

One recent development may strengthen this point. Within the last five years, the digital transformation has reached German municipalities, and has significant impact on all citizens regardless of their social status or age group. "Smart Cities" start dealing with data from many different sources, and citizens play an important role in supplying that data.

By now we can identify countless (inter-)national flagship projects and new technologies; however, most of them are of a rather experimental than a standard-setting character. Examples for large and/or expensive projects can be found in the "Polisdigitocracy" report (Cosgrave et al. 2015) or the Smart Cities Infosystem of the EU (http://smartcities-infosystem.eu/). Smaller projects, often co-created or co-executed by citizens, are documented in the Smart City Charta (BBSR 2017) and on platforms like Challenge.gov or Innocentive.com. What we miss is the discussion about failed projects, including reflections on factors for success or failure, sustainability of the projects, and the risk and mitigation of digital gap in civic society.

According to the norm control board (*Normenkontrollrat*), digitization in German local governments still happens too slowly (DIN e.V. 2017). The potential of big data sources is rated

[1] This famous "quote" is in fact a paraphrase by Samuel S. Wilks; the original (less catchy) quote is as follows: "The time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex worldwide States that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write." (Wells, H. G. 1903)

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

high, but still it is hardly used, even though possibilities are enormous with respect to improvement of citizen service, urban planning, and encouragement of civic engagement. On the other hand, we face thresholds such as privacy issues, proprietary data that is not accessible without restrictions, and finally a lack of data science competence. The latter includes theoretical and practical knowledge in statistics or analytics, but also data management and expertise in the interpretation of the results.

CASE STUDY: THE FRAPORT SMART DATA LAB AT FRANKFURT AIRPORT

But how can we promote the development of competencies for the handling of new media and large, new data sources in civil society, and find ways to make new sources of data as fruitful as possible? We could take a look at companies and their concepts for data-based innovation labs. As an operating company at Frankfurt Airport, Fraport AG has conducted such labs once or twice a year since March 2015. They deal with relevant questions from various departments of the company, e.g., with the improvement of business forecasts.

Digitization provides data about passengers, visitors, shops, transactions, flights, freight and so on—almost anything can be tracked digitally, and the data generated can be used to generate knowledge and value. To create value, these data must be harvested by reduction and abstraction ("How can we track meaningful data?"), cleaned and linked through processing and organization ("How can we generate information from data?"), analyzed, interpreted ("How can we gain knowledge from information?") and applied ("How should we act, based on that knowledge?"). Data "form the base or bedrock of a knowledge pyramid" (Kitchin, R. 2014).

Fraport used a new concept to identify potential improvements and possibilities for optimization based on data analysis—the "Smart Data Lab." This term is widely used to characterize a type of innovation lab that uses data as resources and develops new business ideas from that data, just like the example we will describe now.

A big share of the profit in airport business is realized in retail and properties. An airport mostly acts as concessionaire and not as shop owner. In that way, it earns money based on revenue participation and store rent with high margin and low cost. At Frankfurt Airport, the shopping range is significantly different within the terminals. Product offerings, premium brand availability and shop presentation can vary considerably. At an airport like Frankfurt, up to 200,000 passengers are moving through the terminal infrastructure, passing shops and marketplaces constantly. At the end, the purchase trigger depends on various factors such as personality traits, the route to the gate, shops passed on the way, waiting and walking time, or travel occasion. These are just a few examples of potential influencing variables explaining shopping behavior at an airport; see Figure 1.

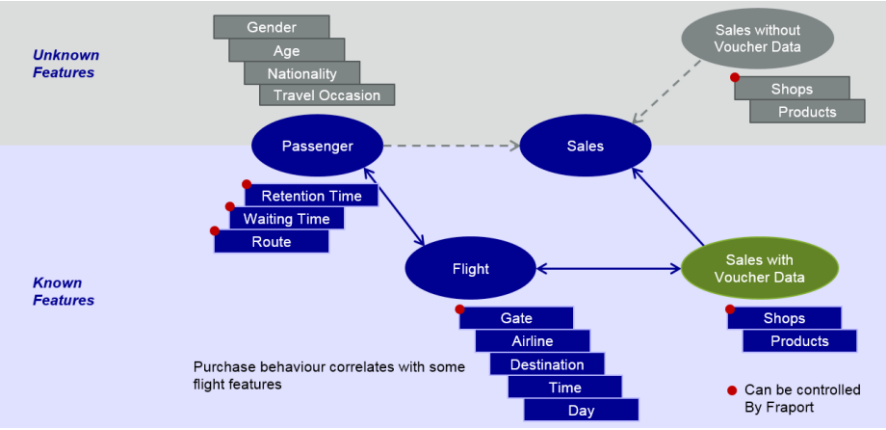


Figure 1: Variables and their relationships in the positioning model.

The positioning of a flight directly affects some of these variables. The task of the Smart Data Lab was to identify important variables and the strength of their influence on retail revenues and purchase probability by combining all available data and processing them with data analytics

techniques. It is important to know which variables can be controlled by Fraport and how to adjust them to maximize retail revenues without running into other operational problems.

An analytical challenge was the substitution of known effects, which was necessary for different reasons. Some of the passenger characteristics cannot be tracked for reasons of data privacy, e.g., passenger nationality. We used destination instead but found that data about the nationality itself, theoretically available from existing data sources, would lead to more accurate predictions. Other characteristics such as spending capacity and the motivation for traveling simply are not available on a customer level. An organizational challenge was the overcoming of existing decision rules that turned out to be myths. It was hard work to explain the difference between correlation and causation, including the issue of spurious or partial correlation, and the difference between a multivariate prescriptive model and descriptive, bivariate analysis. That challenge required excellent communication skills including an idea of political structures among the organization.

Two very important applications can be derived from the results generated out of the Smart Data Lab. First, the insights can be used for future negotiations with airlines about positioning scenarios or the exclusive and prioritized use of terminal areas. Second, current flight planning process, in agreement with the existing arrangement, can be optimized to achieve some quick wins on revenue lift without impairing critical constraints such as guaranteed transfer time.

Figure 2 shows a prototypical 10-week schedule for such a data lab.

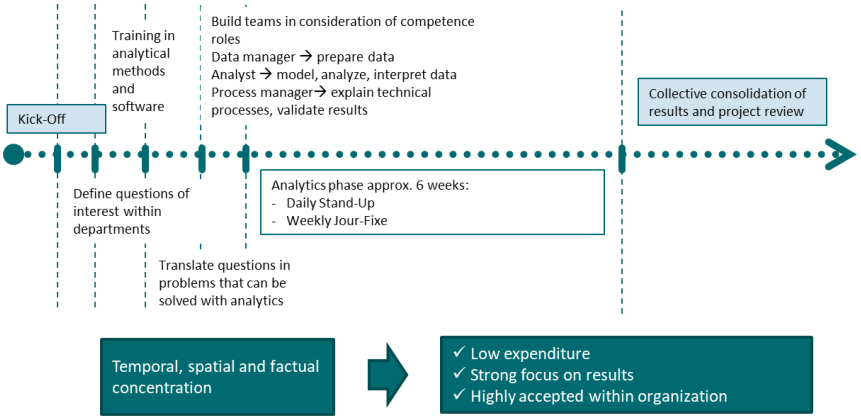


Figure 2: Prototypical schedule of a Smart Data Lab.

TEACHING STATISTICAL LITERACY (1): FROM MOBILE LABS TO HACKATHONS

The possibilities on how to work on data seem endless nowadays, but every analyst needs to start somewhere. To give students the option to try working on data and developing relevant competencies, the following suggestions could be applied as less formal, maybe even extracurricular activities.

Our first suggestion is the concept of Civic Data Mobile Labs. These are low-threshold join-in exhibitions where participants can play with digital data-generating equipment or data in public places. These workshops could be held openly and for a couple of days (one to three days would be appropriate), however, participants may attend for only a few hours or even minutes. Expert guides explain the data and the tools, and may show some previously prepared analyses. Participants will build up analytical thinking and develop their own ideas on what can be done with data.

Another suggestion would be the concept of Digital Workshops. Digital Workshops are multi-method intense discussion workshops with experiments, where interaction and diversity should be highlighted. They can be held in educational institutions like libraries for about one to three days. Their main focus should be on the possibility to learn something from the workshop and to build up and deepen data/digital knowledge. They also offer the chance to take home something self-made, like a nice visualization or even an app.

A third concept that is already frequently used by institutions and companies are hackathons. A hackathon is a workshop in which groups of teams get data and a task to analyze that data. They work on it for a couple of days. Hackathons offer the chance to develop tools,

applications and strategies, accompanied and guided by experts. They are usually held for around three days, which gives participants some time to work deeply through the data, develop their strategies, and try them out. At the end of these hackathons, there is usually a final presentation of the results. There could be prizes for really special tools or insights that help the company (or the municipality), which provided the data, in their future research or development.

TEACHING STATISTICAL LITERACY (2): DATA LABS AT SCHOOL

The city of Moers, the University Rhein-Waal and the Gymnasium Adolfinum started an open data cooperation in 2013, including a variety of data projects. Students in 9th grade were given access to a platform with data about the election results and they worked on questions like "How do votes for the political parties differ among urban districts?" "How do votes differ between elections for Bundestag and Landtag (Federal state elections)?" or "Do election results depend on social structures?" While working on the data, students gave feedback and suggestions on what additional information could be provided for the analysis, such as the age structure of the voters, migration status, working profession, household size, absolute numbers of voters per location, etc.

Another project was completed by Q1st-grade students who analyzed budget data of the City of Moers in Social Science class. They started by analyzing some economical issues in the first half of the year. First they worked on data about statutory duties and voluntary services of the city of Moers, followed by data about the Federal Economic Recovery Plan in Moers, with focus on budget planning in the city of Moers. In the second half of the year they analyzed data about social change. The focus was on urban districts and working places in different economic sectors.

In the upper class for the Q2-grade students, there were data analysis tasks in Social Science class, too. Students worked on municipal data sets mostly from the city of Moers and they worked out some interdisciplinary insights. Students' preferences regarding the problems they worked on were considered wherever possible. Project documentation and communication was done via a project wiki and Twitter (see http://wikifinum.zum.de/wiki/Open-Data).

This case study, as well as our experiences with Fraport Smart Data Labs, inspired us to suggest a novel teaching concept: Smart Data Labs at school.

To start such a novel concept at a school, there is need for a supervising teacher or a team of teachers, students, or data analysts, who plan the lab. They need profound knowledge about data and data analysis. Projects like this need a lot of planning to decide on data sets, software, starting levels, and the degree to which statistical and technical concepts should be explained theoretically.

But apart from the professional questions: How can such a data lab be successful? We believe it is a good idea to teach students that data-driven decision-making includes the willingness to overcome the sugarcoating. We want to learn something new, therefore we cannot know what the outcome will be. Agile methods can be very helpful. They include the use of task boards, which support flexibility (one task per person per day), cooperation between students, and also adaptive planning during the "Daily Standups." Task boards also provide a sort of documentation, which is then aggregated in a project wiki. Weekly *jours fixes* allow students to exchange ideas, discuss them in more detail, and thereby practice a form of peer-teaching. Additional support is given by experts who guide students on realizing their ideas.

The role of the supervisor is more that of a coach than that of a teacher. The whole project should result in a knowledge-building experience for students, in which an analysis expert helps them achieve their goals instead of letting them just experiment and probably fail, because this could result in frustration. Students should try on their own, but when they are stuck with a problem, someone needs to help them. The process itself is the main thing that should be experienced by the students: how to work together, how to develop strategies to solve data-related problems, and how to adapt to the unavoidable challenges. Learning how to use the tools is a side effect.

Exchange and feedback are very important. As described in the Moers example, the students' feedback led to new ideas on what to include in the data and what to do with it. This is fruitful both for students and expert guides. The latter will increase their teaching skills and develop a better understanding of the more data-oriented generations in the future.

Students will have the chance to take home manifestations of their newly acquired knowledge, may it be in the form of charts or analyses which they can be proud of. They are

encouraged to think about data before they act on it. They will develop insights in how data is generated, how data is cleaned and linked to get information out of the data, they learn the basics of how to process, analyze and interpret the data for acquiring this new information, and how this can be used or implemented. No expert is made in a few days or weeks (although some "data science" webinars on the internet might suggest that), but the earlier we start to develop statistical literacy, the greater the chances are that we educate a generation of "Digital Natives" instead of "Digital Naïves."

CONCLUSION

When teaching statistical literacy, it is important to start early. The ability to interpret data correctly will help greatly in every subject of school and university. Statistical literacy is vital to understand the numbers and graphics in all kinds of books, from mathematical and scientific books to social or history books. This is not limited to schoolbooks, but includes newspapers, media and every situation in our daily life when we are confronted by statistics. With the rise of digitization and data-driven decision making, the urge for statistical literacy gets more and more important.

While sitting in Math class in secondary school, one might hear strange words like "averages" and "correlation," maybe even "linear regression." Too often, students are only taught how to work out mathematical solutions for every task they are given, as if there were only one correct and many wrong ways to solve a data task. Sometimes teachers might tell them not to trust every statistical analysis, showing an example of a highly correlated population of stork sightings and newborns and explaining that correlation does not necessarily mean causation.

But to be honest, those widespread examples are not at all related to the life of adolescents. No wonder that they prefer playing with their mobile phones, feeding algorithms somewhere out there that analyze their data. We believe that there is a better way of teaching data science and statistical literacy. Data Labs can be fun as they might address problems from the students' daily lives and show them how they can be solved by the help of data. To learn how to analyze data and how to pay attention while reading statistical analyses are two major skills that can be achieved in Data Labs. Their main advantage is the focus on practically relevant, real-life examples that trigger creativity and bring digitization and data to life.

REFERENCES

BBSR. (2017). *Smart city charta: Digitale Transformation in den Kommunen nachhaltig gestalten*. Bonn, Mai 2017.

Cosgrave, E., Doody, L., & Frost, L. (2015). *Polisdigitocracy: Digital technology, citizen engagement and climate action*. C40/Arup, November 2015

DIN e.V. (2017). *Technologie und Mensch in der Kommune von morgen : Impulspapier zu Normen und Standards.* Berlin, Mai 2017.

Kitchin, R. (2014). *The Data Revolution: Big data, open data, data infrastructures & their consequences*, London: Sage.

Schüller, K., & Förster, A. (2017). Digital Literacy für die Stadt. *Informationen zur Raumentwicklung*, Heft 1/2017, 108–121, und das dort erwähnte Forschungsprojekt Gamification, Prognosemärkte, Wikis & Co.: Neues Wissen für die Stadt

Wells, H. G. (1903). *Mankind in the making*. London: Chapman & Hall.

STATISTICAL EDUCATION IN TIMES OF BIG DATA PERSPECTIVES FROM AN NSI POINT OF VIEW

Markus Zwick
Federal Statistical Office Germany
markus.zwick@destatis.de

INTRODUCTION

With the permanent growth of accessible digital data, commonly denoted as *big data*, the necessary competencies of data producers as well as data analysts are changing. The future competence profile will be different, and the increase of job offers for Data Scientists as well as iStatisticians (informatics statisticians) shows that statistical education also has to develop further (Ridgway 2015).

For official data producers, questions of human resources are essential for the future (UNECE 2013). First of all, National Statistical Institutes (NSIs) need more skills in data science, in combination with other experiences, to produce official statistics of high quality. NSIs are also in stiff competition with companies such as Google or Amazon for the 'best brains.' Well-educated academics with empirical backgrounds are highly sought after, and the price for such resources will increase even further. NSIs will need to be creative in order to acquire the next generation of statisticians. One solution could be to work together with universities on academic programs. The European Master of Official Statistics (EMOS) is one answer to these challenges.[1]

Cooperation between NSIs and universities allows NSIs to influence the curriculum of university programs. It has to be clear what the necessary skills are, and what kind of personnel structure NSIs will have in the future. Most NSIs in Europe have a mix of staff members with master's and bachelor's degrees, coming from different academic fields. In particular, those with bachelor's degrees have often had only introductory courses in statistics at university. As such, master's programs should include more aspects of new digital data sources, and introductory courses in statistics will also need to be further developed. This is the case also for permanent internal training inside the NSIs.

Introductory courses and sometimes master's courses in statistics reach students who may not necessarily work later as a data producer; many may well be on the side of the data users. Introductory courses should be the focus for statistical literacy programs run by NSIs (Forbes et al. 2011).

WHAT WILL BE THE SKILLS FOR THE FUTURE STATISTICIAN?

The *data scientist* is the superstar of the big data age (Davenport et al. 2012), for he or she is able to solve all challenges coming from a large amount of data to produce insights. The data scientist has mathematical and statistical skills, works with a lot of different programs, is able to organize and mix data and metadata, visualize the results in a unique way, and lead teams, as well as write articles in journals.

The data scientist will be part of a team with specialized member skills. The UNECE High-Level Group for the Modernization of Official Statistics has collected a set of team skills necessary to produce official statistics based on new digital data sources as well as a competency profile for big data team leaders (Vale 2016).

The identified big data team level competencies are:

- Interpersonal and communication skills
- Delivery of results
- Innovation and contextual awareness
- Specialist knowledge and expertise

[1] https://ec.europa.eu/eurostat/cros/content/emos_en

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

- Statistical/IT skills
- Data analytical/visualisation skills

The identified big data team leader level competencies are:

- Leadership and strategic direction
- Judgment and decision-making
- Management and delivery of results
- Building relationships and communication
- Specialist knowledge and expertise
- Statistical/IT skills
- Data analytical/visualisation skills

One answer as to how these competencies could be taught derives from the EMOS learning outcomes.[2] EMOS learning outcomes are an up-to-date benchmark for all knowledge content necessary for the next generation of official statistician. During the conception phase of EMOS, professionals from NSIs and universities were able to develop a set of learning outcomes in line with EMOS master's programs. This content was developed with the aim of promoting the right statistical skills, and will be continually subject to development in line with needs.

If new digital data sources (big data as well as administrative data) are best to process in teams, two questions are to be answered. First, how we can teach teamwork and interdisciplinary approaches, and secondly, how prepared are we—inside the NSIs—to work in these kind of teams and produce official statistics?

EDUCATIONAL STRATEGIES FOR NSIS

The face of official statistics will change notably over the next decade. The data landscape has already started to change drastically and the production process, as well as the products of official statistics, will be next.

In order to steer this process, well-considered concepts will be necessary. One part of the strategy aspect is statistical education, on two sides: educating data producers as well as data users. This should include the whole process of education, beginning early in schools. Statistical literacy starts with the pupils. Census at School is a very good example for that.[3] Statistics at school has to be more than probability theory in mathematics courses.

What we need are tailor-made products on official statistics for teacher and pupils, statistical material for economics, geography and/or biology courses. One very good example is 'Bringing Data to Life in the Classroom' of the Australian Bureau of Statistics.[4]

As far as cooperation with universities is concerned, EMOS is a concrete step in the right direction and yet, for the reasons mentioned above, not enough. Consequently, the EMOS idea should cover introductory statistics courses all the way through PhD programs. EMOS is not only an educational program; it is also a network of universities and data producers who work closely together in matters concerning official statistics. The network could and should be used to influence more than the content of master's programs.

We view EMOS as a vehicle for ongoing training inside the NSIs in the future. In particular, in such fast-changing times, permanent training programs inside NSIs and the European Statistical System (ESS) are of utmost importance.

There are currently three different internal training offerings inside the ESS, and they are not yet harmonized. By way of example, each NSI offers courses for their staff in different fields, and there is the European Statistical Training Program (ESTP)[5] as well as EMOS. These single

[2] https://ec.europa.eu/eurostat/cros/content/learning-outcomes_en

[3] http://ww2.amstat.org/CensusAtSchool/

[4] http://www.abs.gov.au/websitedbs/cashome.nsf/Home/Entry%20Page.es

[5] http://ec.europa.eu/eurostat/web/european-statistical-system/training-programme-estp

programs could be more efficient and less cost intensive if we could link them more. An ESS degree for professional statisticians, such as the 'Graduate Statistician' of the Royal Statistical Society, could be one solution.[6]

In all, ESTP, EMOS, and the internal training programs offer enough courses for an ESS degree for a professional statistician; however it would be necessary to develop one or more curricula for an ESS degree. In this way, three independent programs (ESTP, EMOS and internal training) would follow a more harmonized direction.

REFERENCES

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review*, 0, 0, 1–22. doi:10.1111/insr.12110.

United Nations Economic Commission for Europe (UNECE). (2013). *Human resources management and training: Compilation of good practices in statistical offices*. Retrieved 2018 February 13 from http://www.unece.org/fileadmin/DAM/stats/publications/HRMT_w_cover_resized.pdf

Forbes S., Camden, M., Pihama, N., Bucknall, P., & Pfannkuch, M. (2011). Official Statistics and statistical literacy: They need each other, *Statistical Journal of the IAOS, 27*, p. 113 ff.

Davenport, T. H. & Patil, D. J. (2012). Data Scientist: The sexiest job of the 21st century, *Harvard Business Review*, October 2012. Retrieved 2018 February 13 from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Vale, S. (2016). Competency Profiles. [wiki] UNECE High-Level Group for the Modernisation of Official Statistics. Retrieved 2018 February 13 from http://www1.unece.org/stat/platform/display/bigdata/Competency+Profiles

6 http://www.rss.org.uk/RSS/pro_dev/pro_awards/gradstat/RSS/pro_dev/pro_awards/Graduate_statistician/Graduate_Statistician.aspx?hkey=3751895f-02e8-4359-a20b-2a8548cca371

5 Data science initiatives at school level

IF YOU'RE NOT PAYING FOR IT YOU ARE THE PRODUCT: A LESSON SERIES ON DATA, PROFILES, AND DEMOCRACY

Bettina Berendt, Gebhard Dettmar
KU Leuven, Department of Computer Science, Celestijnenlaan 200A, 3001 Heverlee, Belgium
Behörde für Schule und Berufsbildung, Hamburger Straße 31, 22083 Hamburg
bettina.berendt@cs.kuleuven.be, g.dettmar@web.de

We developed and performed (in school) a lesson series that covers the effects of internet tracking and data analysis by the data collecting industry—Facebook, Google & Co.—on a political system that ranks autonomy and privacy among its basic rights.

Various court judgments, including, in Germany, the Lüth judgment by the German Constitutional Court (BVerfG), have established that these rights are positive fundamental rights that are indispensable for political participation, and that therefore bind not only state actors, but also actors such as companies.

Accordingly, the series introduces, step by step, data protection related issues such as tracking and targeted advertising, data mining, and social exclusion (health insurance / credit rating as examples), as well as psychometry applied to Facebook likes ("Big 5 Personal Traits" plus IQ). Based on this, the series deals with the consequences for privacy, their function for the possibilities of participation within democracy, and the legal guarantees attached to it by the BVerfG.

We report on the genesis, conceptualization, and experience with the implementation of this series, and will further discuss developments since the first iterations of the lesson series and sketch how these could be integrated, in particular the debate around filter bubbles and fake news (on Facebook and other social media).

THE LESSON SERIES IN DETAIL

The lesson series starts with the visualization of internet tracking via browser plugins, thus revealing the core business of the so-called data leech industry: data collection, data evaluation, and profile creation. These profiles are a valuable product which is sold to other companies. A data collection company has to inform its users about this in a privacy policy which is a part of the terms of service.

Here is an excerpt from Facebook's data use policy, status quo 2013[1]: "We do not share any of your information with advertisers (unless, of course, you give us permission). As described in this policy, we may share your information when we have removed from it anything that personally identifies you or combined it with other information so that it no longer personally identifies you.[2] […] We use the information we receive, including the information you provide at registration or add to your account or timeline, to deliver ads and to make them more relevant to you. This includes all of the things you share and do on Facebook, such as the Pages you like or key words from your stories, and the things we infer from your use of Facebook."

The striking term in this excerpt is "things we infer from your use of Facebook." The main purpose of this lesson series is to give a deeper understanding of the potential of machine learning algorithms to infer things from a user's use of Facebook.

The social implications of data aggregation performed by machine learning algorithms are aptly described in an article by Lori Andrews in the New York Times. She reports the case of an Atlanta man returning from his honeymoon to find his credit limit lowered from $10,800 to $3,800. A letter told him: "Other customers who have used their cards at establishments where you recently shopped have a poor repayment history with American Express." The establishment he recently shopped in was a guitar shop. A term that was popular in the 1970s, "redlining," is thus changing to "weblining." In the author's words: "The term *weblining* describes the practice of denying

[1] Data Use Policy, https://www.facebook.com/about/privacy/ retrieved 06/30/2013
[2] To the issue Pseudoanonymizing personal data see Berendt et al. (2015)

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

people opportunities based on their digital selves. You might be refused health insurance based on a Google search you did about a medical condition. You might be shown a credit card with a lower credit limit, not because of your credit history, but because of your race, sex, or ZIP code, or the types of Web sites you visit."[3]

It is already clear that decisions entirely based on correlation are, from a user's point of view, totally non-transparent and beyond comprehension.

At this stage of the series it is necessary to give further insight and hands-on experience in the working methods of machine learning algorithms. We chose the Apriori algorithm for frequent itemset mining; and as a database, we took a small sample of Facebook's Graph API.

User ID	Favorite athletes (field)	Education	Relationship Status
1	Basket Ball	Gutenberg-Gymnasium	+
2	Soccer Player	Helmut Schmidt-Gymnasium (HSG)	-
3	Soccer Player	Helmut Schmidt-Gymnasium	-
4	Soccer Player	Gutenberg-Gymnasium	+
5	Soccer Player	Helmut Schmidt-Gymnasium	-

The resulting association rules (candidate frequent 2&3-itemsets) are:

Soccer -> HSG: support = 60%, confidence = 75%
Gutenberg-Gymnasium -> +: support = 40%, confidence = 100%
Soccer & HSG -> -: support = 60%, confidence = 100%

While the last rule in particular created a lot of amusement in the courses we taught at HSG, the rules also raised suspicion: Is it really possible to conclude so much, and with complete (100%) certainty, from data about a person that is totally unrelated to their romantic gifts? Obviously, this tiny example is a toy example, and the rules over-generalize starkly. In the subsequent part of the lesson series, we therefore turned to a larger and real-life example.

In 2013, the Psychometric Centre of the University of Cambridge announced that "easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender."[4] In fact, the Psychometric Centre tried to establish a tool, "designed for use by brands and agencies," PreferenceTool, that uses Facebook Likes to determine the big 5 personal traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism, or OCEAN) and intelligence (IQ). The meaning of the several attributes is described like this: "IQ […] is used extensively both in educational settings, to distinguish individuals with a high ability from those who will need more help; and in work settings, to hire and promote employees. It is the best predictor of job success."[5]

There are two Facebook Likes which are extremely popular among students: Converse and Dunkin Donuts. The results for IQ and Conscientiousness, the most important attributes in education and work settings, are disastrous: People who liked these two items are "more intelligent than 25% of the population, putting them in the very low range." Students matching a result like this are obviously not suited for higher education and a promising job career. The students in the courses we taught reacted with great indignation to this aspersion.

The completely non-transparent character of purely correlational profiling and its implications for privacy have been made clear so far. We now deal with the function of privacy within democracy and its legal consequences.

[3] Andrews (2012), http://www.nytimes.com/2012/02/05/opinion/sunday/Facebook-is-using-you.html
[4] Kosinski (2013), p.1.
[5] see http://preferencetool.com/ for all attributes. Kosinski et al.'s ideas and app design have had a later re-incarnation in the context of the work of Cambridge Analytica and its alleged meddling in the 2016 US presidential election.

In 1983 the Constitutional Court of Germany defined informational self-determination as a fundamental right in the "Census judgment." The government had planned a comprehensive census because it considered the to-be-collected statistics necessary for shaping essential policies concerning, for example, markets, health and housing. The Court ruled against this census mainly because of the planned processing of personal information and its linking with population registers, resulting in a data ensemble that would not be anonymized and that would contain rich profiles of individuals and groups. The possibility of data linking and uncontrolled personality profiling was considered a threat to privacy and democratic participation. Since that time, informational self-determination has been a basic pillar of German and European data protection law, which constrains public and private actors. The pioneering ruling contains a fundamental description of privacy's function within democracy: "[The data] can also be combined—in particular, when integrated information systems are set up—with other data collections into a partial or complete personality model, without the individual being able to control this model's correctness or use. [...]. The right to informational self-determination would be impossible to exercise in a societal order and underlying legal order in which citizens can no longer know who knows what, when, and in which context, about them. People who are uncertain about whether their divergent behavior is continuously registered and persistently stored, used, and transferred as information, will try not to draw attention to themselves through such behaviors."[6]

Since it is has been made clear so far that methods predicting attributes like IQ by evaluating Facebook Likes are incompatible with the right of informational self-determination, the question is how a careful study of Facebook's Data Use Policy is at all sufficient for giving a so called "informed consent," and whether giving an informed consent can be sufficient for the abandonment of fundamental rights. In German context the Lüth judgment is relevant here: it defines the fundamental rights as not only negative rights ("defenses") of the citizen against government, but also as positive rights ("claims"), therefore as indispensable and inalienable. They have an emanation effect, an indirect third effect ("mittelbare Drittwirkung") that emanates into all areas of law. Thus, since the freedom of contract in private law is affected, this emanation binds private actors. There have been objections to this point of view.

We discuss these objections at the end of the lesson series in a role play: a Facebook user complains about FB's profiling of his user account. Both freedom of contract and right to data protection are fundamental rights. So should the first right be curtailed in order to protect the second right, or would such curtailing amount to an overprotective "nanny state" that in the end harms citizens' autonomy? Historically, these two positions have been debated fiercely, and they re-surface periodically. They were crystallized into the following two positions, which we also encouraged students to assume in the lesson series:

1. The position of the Constitutional Court is: Privacy-invading private contracts imply a suspension of fundamental rights, thus forcing the citizen to abdicate his right of, e.g., freedom of speech; this is an example of the so called chilling effect, which is unconstitutional.
2. The position of jurists such as Böckenförde is: the self-responsible citizen is incapacitated by an interpretation of fundamental rights that transforms the Constitutional Court of Germany into some kind of Constitution-Areopagus that dictates societal values in form of a value-oriented constitution.

The discussion touches an important debate in jurisprudence: the problem of freedom among unequal citizens which has been described (by Garaudy) as "the freedom of a free fox in a free henhouse."

CONCLUSION

As a result of this lesson series, the students have to further discuss the value of privacy in a society that increasingly ignores these rights.

[6] See full text German version here: https://openjur.de/u/268440.html

A desideratum in this lesson series is to understand the implications of algorithmic decision-making with regard to: a) biased data, echo chambers, filter bubbles, fake news and "digital mass persuasion"; and b) big data for monitoring educational systems with consideration of students' privacy, educational equity and efficiency, student tracking, assessment, and skills.

REFERENCES

Berendt, B., Littlejohn, A., Kern, P., et al. (2017). *Big data for monitoring educational systems*. Luxembourg: Publications Office of the European Union.

Berendt, B., Dettmar G., et al. (2016). Datenschutz im 21. Jahrhundert—ist Schutz der Privatsphäre (noch) möglich? In J. Gallenbacher (Ed.), *Informatik allgemeinbildend begreifen*. INFOS 2015 (pp. 33–42). Bonn: Gesellschaft für Informatik, *Lecture notes in informatics* (LNI). http://www.infos15.de/GI_Proceedings_Band-249_incl.pdf

Berendt, B. & Coudert, F. (2015). Privatsphäre und Datenschutz lehren - Ein interdisziplinärer Ansatz. Konzept, Umsetzung, Schlussfolgerungen und Perspektiven. [Teaching privacy and data protection - an interdisciplinary approach. Concept, implementation, conclusions and perspectives.] In *Neues Handbuch Hochschullehre*. [*New Handbook of Teaching in Higher Education*] (EG 71, 2015, E1.9, 7–40).

Berendt, B., Dettmar, G., Demir, C., Peetz, T. (2014). Kostenlos ist nicht kostenfrei oder: If you're not paying for it, you are the product. *LOG IN*, 178/179, 41–56. http://people.cs.kuleuven.be/%7Ebettina.berendt/Papers/berendt_dettmar_demir_peetz_2014.pdf

Kosinski, M., Stillwell, D. and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* (PNAS). http://www.pnas.org/content/110/15/5802.full

Andrews, L. (2012): Facebook is using you. *The New York Times*, Feb. 4, 2012, http://www.nytimes.com/2012/02/05/opinion/sunday/Facebook-is-using-you.html.

Dreier, H. (1993). Dimensionen der Grundrechte. Von der Wertordnungsjudikatur zu den objektiv-rechtlichen Grundrechtsgehalten (*Schriftenreihe der Juristischen Studiengesellschaft-Heft 23*), 9–64.

LEARNING FROM DATA ABOUT SOCIETY: PERSPECTIVES AND EXPERIENCES FROM PROCIVICSTAT

Joachim Engel
Ludwigsburg University of Education, Ludwigsburg, Germany
engel@ph-ludwigsburg.de

Data are abundant. Quantitative information about the state of society and the wider world is around us more than ever. Paradoxically, recent trends in public discourse point towards a post-factual world that seems content to ignore or misrepresent empirical evidence. As statistics educators, we are challenged to promote understanding of statistics about society. In order to re-root public debate to be based on facts instead of emotions and to promote evidence-based policy decisions, statistics education needs to embrace two areas widely neglected in secondary and tertiary education: understanding of multivariate phenomena and the thinking with and learning from complex data.

BACKGROUND

Digital media and the availability of data of sheer unlimited scope and magnitude change our access to information in radical ways. Emerging data sources provide new sorts of evidence, provoke new sorts of questions, make possible new sorts of answers, and shape the ways that evidence is used to influence decision-making in private, professional and public life. In addition, burgeoning progress in hardware and software development and the inception of computer intensive algorithms pave the way to powerful methods of data analysis, from innovative visualizations of complex data to methods of data mining and machine learning. These advances are changing the nature of the evidence that is available, the way it is presented and used, and the skills needed for its interpretation (Ridgway 2015). Recent developments impact not only the statistical methodology and empirical methods of scientific inquiry, but affect society and its citizenry as a whole. Challenges for statistics educators include the communication of concepts and ideas to empirical researchers in other fields, but also the challenge to enable the public to understand and reason about quantitative evidence in a world awash with data. Bill Cleveland, acclaimed for being the first person to use the term "data science" (Cleveland 2001), emphasized the need to pay attention to pedagogy and pleaded for academic statistics departments to devote energy and resources to education.

Data on important societal topics are becoming increasingly accessible to the general public and to individual citizens or social action groups, on a huge range of topics such as migration, employment, social (in-)equality, demographic changes, crime, poverty, access to services, energy usage, living conditions, health and nutrition, education, human rights, and many others. In order to ground public debate in facts instead of emotions and to promote evidence-based policy decisions, statistics education needs to embrace two areas widely neglected in secondary and tertiary education: understanding of multivariate phenomena and the thinking with and learning from complex data (Engel 2016). This paper emphasizes the need for innovation in statistics and computer science curricula at the secondary and tertiary educational level to prepare students to become statistically and digitally literate citizens; and discusses the role of data science education as an indispensable ingredient in the preparation of young people to become informed and engaged citizens in the digital age.

CIVIC STATISTICS AND DATA SCIENCE EDUCATION

In an increasingly complex world, the involvement of informed and committed citizens is a critical resource in public decision-making at international, national, and local levels. The project ProCivicStat, a strategic partnership of six universities funded through the Erasmus+ program of the European Union (funding period September 2015 to August 2018), explores a subfield we call *Civic Statistics*, which focuses on understanding quantitative and statistical information about society as provided by the media, statistics offices, and other statistics providers (Engel et al. 2016). Understanding Civic Statistics is required for participation in democratic societies, but involves data that often are open, official, multivariate in nature, and/or dynamic, that is not

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

normally taught in regular mathematics and statistics education, let alone in civics or social studies classes. Few high school teachers in mathematics receive any training on how to teach statistics, not to speak of social science teachers who may have no training in statistics at all. As a result, teachers stay within their comfort zones and overemphasize a narrow range of statistical techniques and computations (mathematics) or fail to engage with statistical ideas at all (social science). They pay too little attention to working with and understanding the multivariate data that describe social trends, and to the analysis, interpretation and communication about the meaning of such data (Cobb 2015). But capacity building for informed and committed citizens has to start in school education. While focusing on curricula at the secondary and tertiary level, the ultimate goal of ProCivicStat is to strengthen civil society, empowering informed citizens for evidence-based decision-making and civil society engagement. The challenge is multi-faceted. Data literacy for civic engagement involves, among many other aspects, specific statistical knowledge, ICT skills, knowledge about computing and data structures, critical thinking, and much more.

William Finzer (2013), the developer of the data science education platform CODAP (http://codap.concord.org/), emphasizes important "data habits of mind" as a collection of attitudes and reflexive approaches to understanding the world through data

- Acknowledge the need for data to gain insight.
- Look for data: Ask "Which data might be helpful to reach conclusions, gain insight or build arguments?"
- Graph the data: Construct graphical representations that highlight potentially useful patterns in the data, patterns that are difficult to discern by staring at a table of numbers.
- Ask new questions, explore relationships, and search for answers using different visualizations and calculations.
- Become immersed in the data: Use (and invent new) measures and parameters. Explore and search for the story behind the data.
- Ask: "Why do the data vary? Are there systematic reasons or are the fluctuations more random? Are there other factors that can convey or explain the relationship between two variables?"

From the perspective of dealing with complex data about society we add

- Search for third, explanatory variables that may be related to an observed relationship between two variables.
- Develop a critical view on the quality and source of the data. Question the politics of the data. How were the data collected? How were the variables defined? How were constructs operationalized? Why, with what purpose, and in whose interest were the data collected?

Students should know that data and information from large surveys are available from recognized organizations and can be used to gain new insights. If they see a statistic in the news, on social media, or on a website, they should know how to check the source and decide how trustworthy the reported information is. Blind, uncritical trust in data is no less dangerous than an *a priori* rejection of evidence-based thinking on the basis of distrust in any statistic.

Initiatives such as data.gov in the USA and data.gov.uk in the UK aim to support the democratic process by giving citizens access to data that can stimulate debate and inform policy making—however, accessing and working directly with such data sets often requires considerable technical expertise. Examples include data from wearable devices, transactional data from mobile phones, and data scraped from web pages. Civic Statistics requires an understanding of the analytic techniques suited to accessing and analyzing high-volume unstructured data, including cleaning and restructuring data. ICT skills are required to engage with ICT-based tools such as statistics packages and knowledge of how to search for information in the World Wide Web. For Civic Statistics, students must use interactive displays effectively.

EXAMPLES FROM PROCIVICSTAT

In ProCivicStat, in addition to conceptual blueprints for understanding multivariate phenomena in a data-rich world, suitable dynamic visualization tools such as Gapminder (https://www.gapminder.org) or displays of the Smart Centre at Durham University for multivariate data (https://www.dur.ac.uk/smart.centre/) were explored and authentic and relevant data sets provided. A main objective was the development, testing, and evaluation of teaching and learning materials for innovative teaching for a wide range of target groups. Teaching materials, extensive datasets, and conceptual representations of civil statistics are available through the website http://www.procivicstat.org.

At Ludwigsburg, materials for teachers and learners were developed for a broad range of socially-relevant contents and topics. In addition to a content-related introduction, the materials contain references and links to additional information to delve deeper into the content. A multivariate data set containing quantitative information related to the content is made accessible with a description of the data source and the variables involved. A link to an electronic worksheet lets the user process the inquiry tasks with the help of CODAP software. To give some examples, the following topics are presented in the described format (worksheets can be accessed via http://www.procivicstat.org):

- Gender Pay Gap: Why do women earn less than men? (Worksheet in Figure 1,)
- Are referees in European football racially biased?
- Has the number of crimes in Germany increased in the wake of the 2015 influx of refugees?
- Do women with higher education tend to have fewer children?

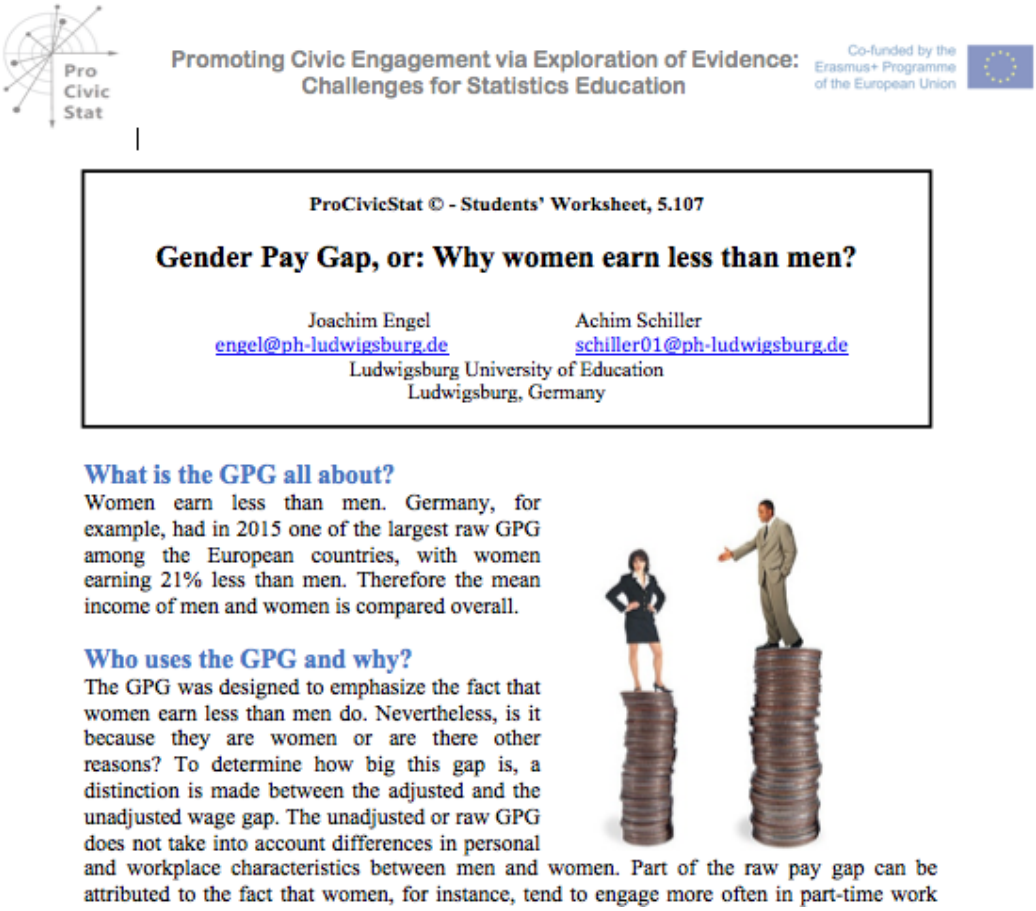


Figure. 1: Head of the worksheet: Why do women earn less than men?

To process these worksheets requires not only basic mathematical and statistical knowledge but also statistical skills that are barely dealt with in traditional lessons—skills such as multivariate thinking, the search for hidden third-party variables and interactions, understanding

correlations and Simpson's Paradox, the exploration of functional relationships between variables and the use of different representations and visualizations. It requires critical thinking: how were variables defined? How were constructs operationalized (e.g., poverty risk or unemployment)? How, why, and by whom were the data collected?

In the process of dealing with these tasks, digital literacy and data science skills are essential to analyze the problems presented and to come to conclusions. Some tasks require searching for more information, including additional data. These data may have to be cleaned, tidied, imported into software, and possibly restructured. Suitable graphical and numerical tools to represent the data have to be chosen. More information may need to be searched for. Depending on the depth of the planned analysis, algorithmic thinking and knowledge about algorithms, e.g. for nonlinear regression or tree-based methods (see below) may be useful. Above all, civic statistics provokes critical evaluation and reflection.

The Common Online Data Analysis Platform (CODAP) proved to be a well-suited environment to process, explore, and analyze civic statistics data. With the help of CODAP, new variables can be defined, and variables can easily be transformed and aggregated. Hierarchically ordered data can be restructured with simple drag-and-drop movements. Data import including web scraping is supported. Investigations and comparison of subgroups are easy. Visualizations, e.g., to compare distributions or to explore how two variables co-vary, are straightforward to do, including curve fitting in scatter plots.

It is not the purpose of this section to present and discuss complete worksheets. The interested reader is referred to the ProCivicStat website (www.procivicstat.org). Rather we illustrate the capacity and challenges in handling civic data with CODAP with a few examples. Figure 2 shows the distribution of hourly pay for men and women separately for the eastern (former GDR) and western part of Germany. There is a striking pay gap only in the west. Notice also from the display that while men in the west have a substantially higher hourly pay than men in the east, there is barely any difference for women between east and west. This observation calls for a search for explanatory lurking variables.

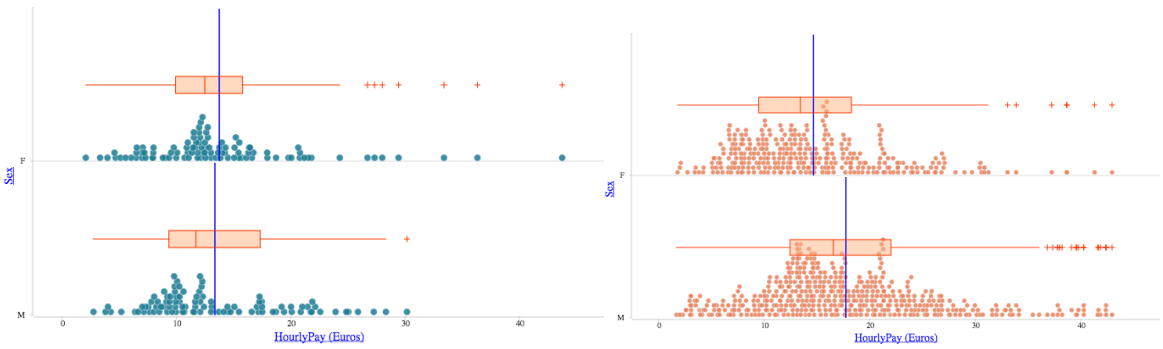


Figure 2: Hourly pay of German employees, separately for the eastern part (former GDR, left) and the western part (right). Notice while in the western part mean (blue line) and median is roughly 3 Euros more for males, there is slight advantage for females in the eastern part). Also there is a substantial pay gap between the eastern and the western part. Data are a random sample (n=1000) of the 2006 income structure survey of the German Statistical Office.

Next, consider data about the living standard in countries around the globe over the last 20 years. These data are available through the World Bank and other international institutions. Figure 3 displays a scatter plot including a fitted curve of the average number of children per woman and the average GDP per person in 229 countries of the world for the year 1984. The continents are color coded in the display. Creation of the graphical display requires structuring the data according to the variables Year (highest level) and Continent (secondary level).

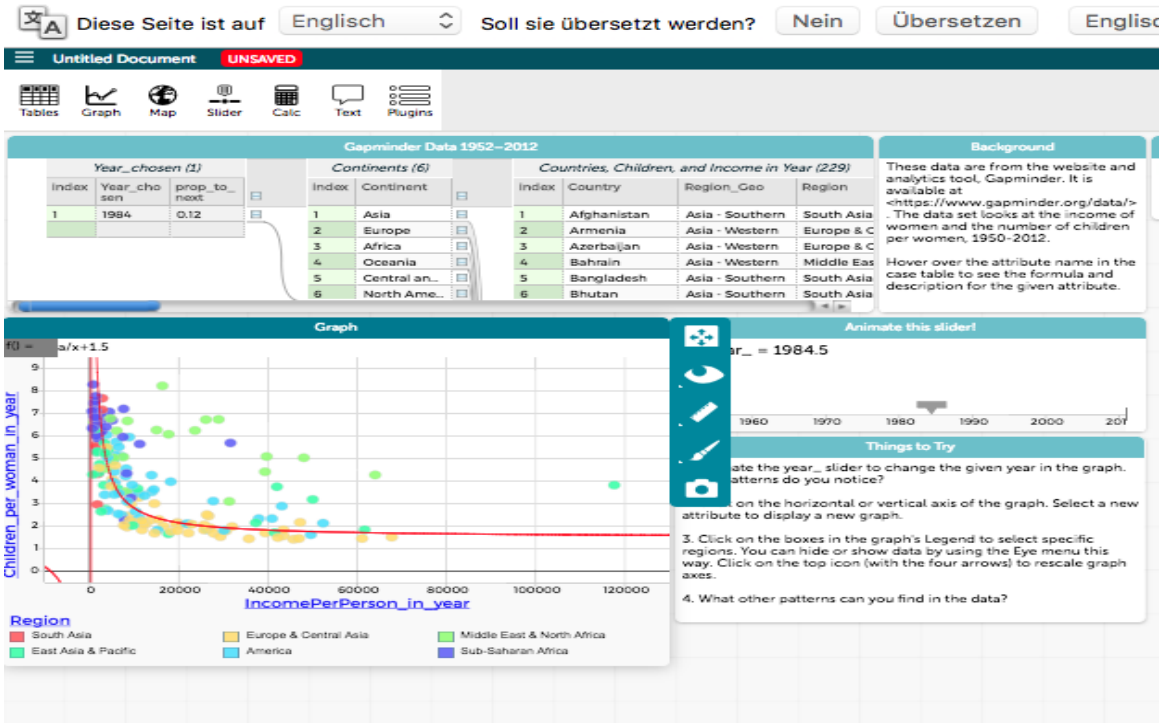


Figure 3: Relation between average number of children per woman and GDP per person. Representaton with CODAP; Each dot represents a country, continents are coded by color.

Another example deals with potential racial bias of referees in European football. More specifically, the question is whether players with dark skin tone are more likely than light skinned players to receive red cards from referees. The data set comprises information on 1419 football players in four professional European football leagues with 19 variables such as number of red, yellow-red and yellow cards received during a player's career, position played, height, weight, and a rating of skin color (1 = very light, 5 = very dark). The data are from a crowdsourcing data analysis project and described in Silverzahn et al. (2014).

While it is easy to create boxplots of the variable RedCard or the derived variable RedCardRate (= red cards per game), the challenging question is the search for possible confounding variables, i.e., third explanatory variables that may account for an observed relationship between the variables RedCardRate and Skintone. Are red cards and skin color equally distributed across the four countries (England, France, Germany, and Spain)? What about the distribution of the position across players of different skin color? Are players of color more often represented in some positions than in others?

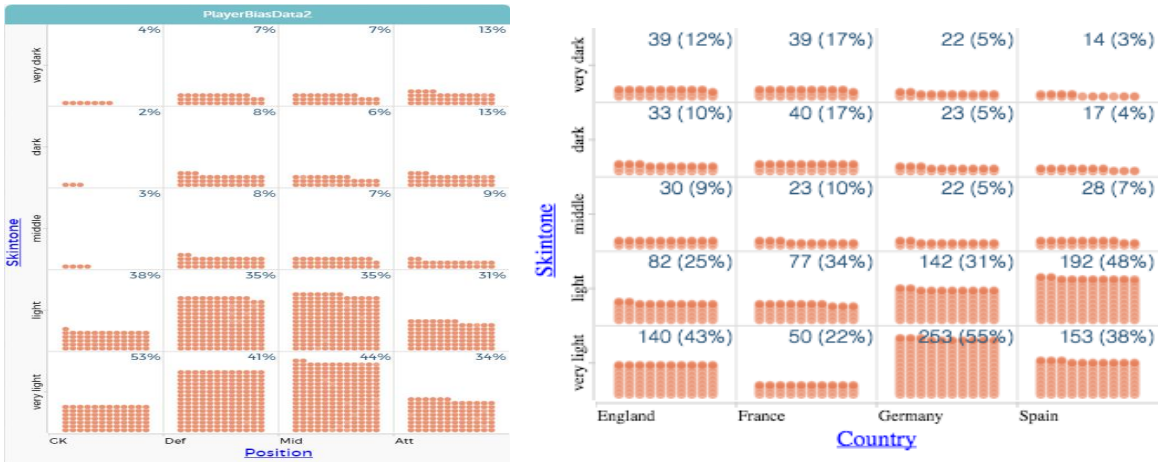


Figure 4: Bivariate distribution of the nominal variables Position and Skintone (left) and Country League and Skintone (right)

Figure 4 displays the distribution of Position versus Skintone and Country League versus Skintone. The figures reveal that dark skinned players are more often attackers or midfielders than defenders or goalkeepers. Also, the French and British Leagues have a higher percentage of darker-skinned players than Germany or France. These observations matter to our guiding question for potential racial bias of referees. For red cards tend to be given more often to defenders than to midfielders or attackers. Also, referees in England, Spain, and France tend to brandish red cards more often than in Germany (Figure 5). All of this calls for caution in jumping to quick conclusions because Country League and Position may well be confounding variables that may have a strong impact on referees' red card giving, thus masking the skin tone variable.

In summary, we conclude by means of the diagrams that darker-skinned players are increasingly attackers. The position of midfielder and defender have nearly the same proportion of all skin tones while darker-skinned goalkeepers are underrepresented. Likewise, the variable RedCardRate is not evenly distributed across the four countries.

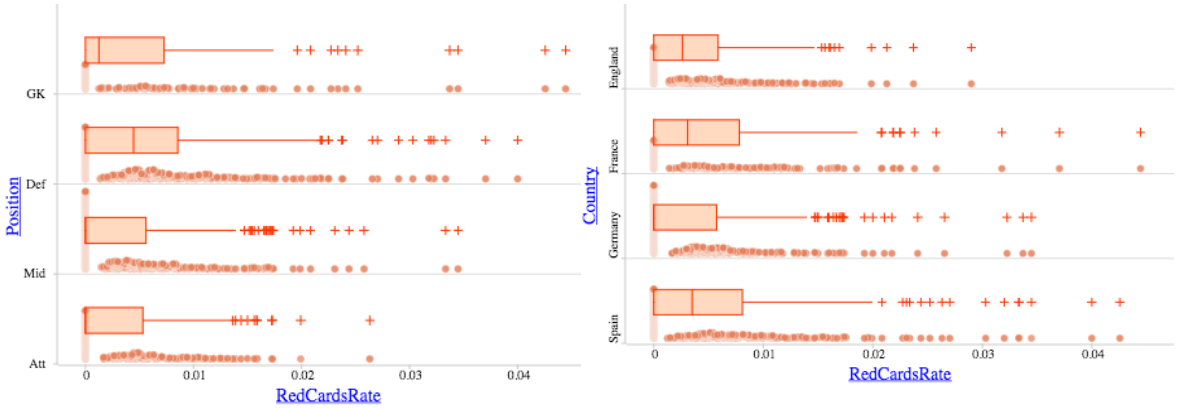


Figure 5: Distribution of RedCardRate across the position of a player (left) and across country leagues (right)

TREE-BASED METHODS FOR LEARNING FROM DATA

Understanding multivariate data is substantially supported by modern statistical methods and modes of presentation. Classification and Regression Trees (CART) are *one* possible method that is particularly suitable for the study of such complex data (Breiman et al. 1984). The interpretation of these trees is fairly straightforward, but the trees are based on computationally very intensive algorithms. These algorithms are the foundation for some deeper statistical learning methods under such sonorous names as bagging, boosting, and random forests (Hastie et al. 2006). With the help of the "Arbol" plug-in, developed in collaboration with Tim Erickson during his 2017 visit to Ludwigsburg, CODAP supports the understanding of how classification and regression trees are constructed, thereby facilitating the interpretation of tree structures. Figure 6 shows a regression tree created with Arbol showing the dependency of the number of red cards per game ("RedCardsRate") in European football on the covariates skin color, league (England, France, Germany, or Spain), and the position of the player. For further details of tree-structured statistical methods, reference should be made to the literature (Breiman et al 1984, Hastie et al., 2006).

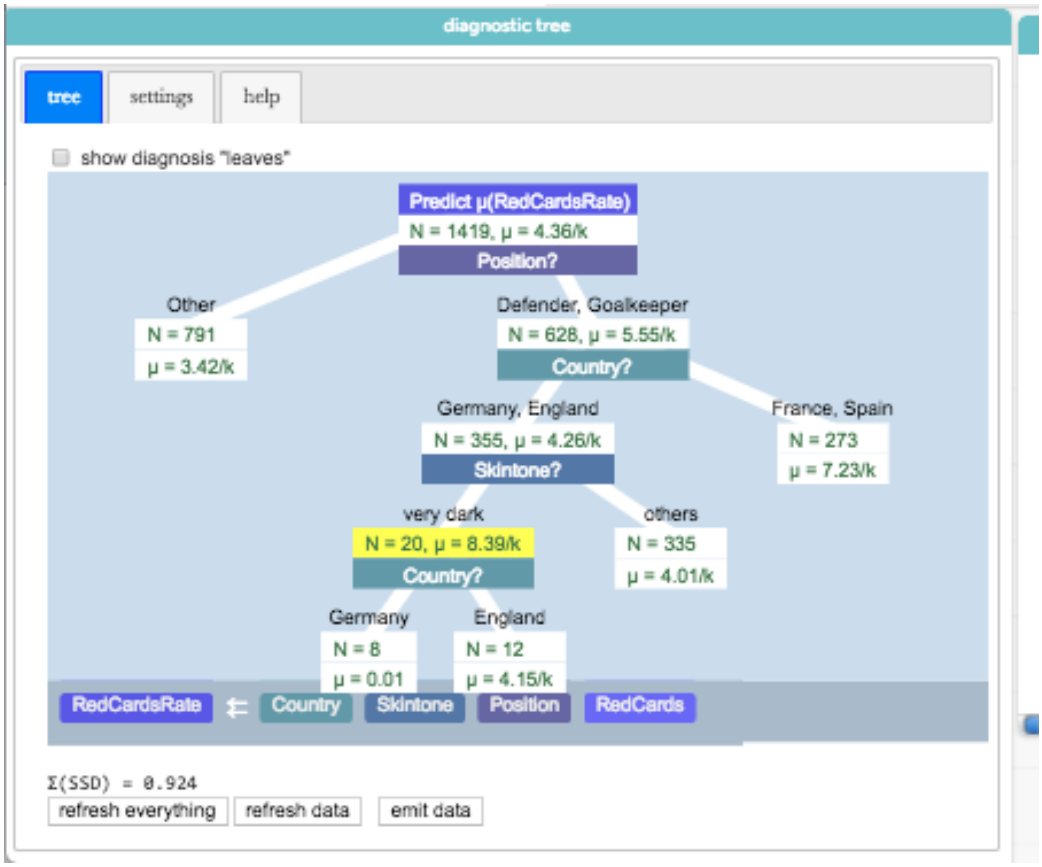


Figure 6: Regression tree for the number of red cards per game depending on the skin color, the player's position, and the country league (created with Arbol, a plug-in tool for CODAP).

CONCLUSION

Recurring statistical topics in our Civic Statistics seminar included (1) comparing distributions (2) aggregating data (3) restructuring data (4) investigating and comparing subgroups (5) measurement and operationalization of variables (6) search for explanatory third variables (7) modeling of functional relationships and (8) inquiring about metadata. Beyond statistical know-how and a willingness to engage with socially relevant topics, digital skills for information search and data management, computational knowledge, and data habits of mind are crucial for making sense and critically understanding the complex data that are typical for Civic Statistics.

ACKNOWLEDGMENT

The work reported in this paper was supported in part by the ProCivicStat project, a strategic partnership of the Universities of Durham, Haifa, Ludwigsburg, Paderborn, Porto, and Szeged, funded by the ERASMUS+ program of the European Commission. However, the views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the funding agency.

REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984). *Classification and regression trees*. Belmont, California: Wadsworth.

Cleveland, W. (2001) Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*, 21–26.

Cobb, G. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician 69*(4): 266–282.

Engel, J. (Ed.) (2016). Promoting understanding of statistics about society. *Proceedings of the IASE Roundtable*. Retrieved from http://www.iase-web.org/Conference_Proceedings.php

Engel, J., Gal, I., & Ridgway, J. (2016). Mathematical literacy and citizen engagement: The role of civic statistics. Paper presented at the 13th International Congress on Mathematics Education (ICME13).

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Reserach Journal (SERJ), 16*(1), 44–49. Retrieved from http://iase-web.org/documents/SERJ/SERJ16(1)_Engel.pdf

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2). escholarship.org/uc/uclastat_cts_tise

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference and prediction* (2. Ed.). New York: Springer.

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/insr.12110/full

Silberzahn, R. Uhlmann, E. L, Martin, D., Nosek, B. et multis al. (2014). *Many analysts, one dataset: Making transparent how variations in analytical choices affect results*. https://osf.io/47tnc/ (retrieved 2016 Feb 4)

THOUGHTS ON DATA SCIENCE EDUCATION AT THE SCHOOL LEVEL

William Finzer, Senior Scientist
Concord Consortium, Emeryville, CA
wfinzer@concord.org

This is a short paper describing my session on "software for learning and doing data science" plus reflections on data science education coming from the Paderborn symposium.

SOFTWARE FOR LEARNING AND DOING DATA SCIENCE

Paraphrasing Biehler (1997) "While software is only one aspect of data science education, the quality of software can have a major impact on the quality of education." And so we must think carefully about the design and capabilities of the software we put in students' hands. But not only that. We need software to help us discover what it means to be learning data science, and, at the same time we need to know what we want students to learn in order to help us think about what learners should be able to do with the software.

GUIDELINES FOR DESIGNING DATA SCIENCE EDUCATION SOFTWARE

The Common Online Data Analysis Platform (CODAP) under development at Concord Consortium, and currently being used by approximately twelve curriculum development projects, is by no means the ideal for students to use for learning and doing data science. Nevertheless we can use it to help stimulate thoughts about what we would like to see in such software. The following list came partly from my thinking and partly from the symposium audience during my presentation.

Get lots of data in easily—It should be trivially easy to get data into the software and there should be many different ways to do it. CODAP makes it easy to import csv data through drag and drop. CODAP plugins make it easy to create front ends to large datasets and databases. Can we also make it easy for students to scrape data from HTML and PDFs without cluttering up the interface and overloading gestures such as dropping things into CODAP document?

Be context neutral—Adapt to any and all data situations so that learners have a tool they can use again and again in different subject areas and over a range of grade levels.

Make data moves natural—Even if they have to be learned rather than discovered, data moves should be easy to remember and make sense. In CODAP, selecting cases in one graph to see the selection in another—or to filter the visible cases— is a successful example of this.

Be structural—To get a taste of data science and to engage in data modeling, students should experience at least one data structure other than flat, row by column. CODAP allows for dynamic hierarchical restructuring and limited join of separate datasets through use of lookup functions.

Handle more than numbers and strings—Data come in a wide variety of types. Experiencing some of these will broaden learners' understanding of what is possible. Go gently, inferring type whenever possible so as not to require uninformed decisions. Beyond strings and numbers, CODAP works with dates, geojson boundaries, and colors, but not yet with images and sounds.

Do GIS—Data are often distributed geographically. Provide tools for visualization on maps. CODAP plots lat-long points and geojson boundaries.

Animate—First, animate transitions so that the learner has a chance to understand how one visualization relates to the next. Second, give the learner the ability to create animations as part of their presentations of results to better convey what they have found. CODAP does a fair job of animating transitions and provides a slider that can be used in formulas to produce user-designed animations.

Be extensible—Rather than attempt to provide all the capabilities users might want, allow the environment to be extended. Drop-in "plugins" extend CODAP's capabilities with data frontends, simulations, games, new kinds of visualizations and analyses, and more.

Automate—Create the ability to pour data in one end and have it come out the other transformed. CODAP does not yet do very much in this regard.

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

Do big data—Make it easy to work with very large datasets without getting bogged down. This is a constant challenge for a browser-based tool such as CODAP. Plugins that serve as an interface to large datasets and that load random samples would at least partially solve the problem.

Create novel visualizations—Part of the fun of working with data comes from inventing interesting displays that bring out relevant features of the data. For now, CODAP is limited to displaying points, so the challenge is to provide more flexibility in data display without making the interface much harder to learn.

Create models—Modeling is as much a part of data science as it is of statistics, and being able to compare data with output from models is essential. Are there types of models that are more commonly used as part of work in data science as opposed to statistics? So far, CODAP is able to produce only the most basic of models, and these are statistical.

Create simulations—In statistics education settings, simulation of the inferential framework has proven powerful in helping learners understand how inference works. In data science we think that simulations that generate data will be helpful to students as they grapple with situations in which it is difficult or impossible to work with actual data. The ease with which plugin simulations can be embedded in CODAP and produce data that learners can analyze lessens the need for a simulation capability that is part of the core learning environment.

Relate—In learning and doing statistics it is relatively rare that a learner must relate two or more datasets as part of the analysis. Not so in data science where even at the learning stage joining related datasets is important and commonplace. Though CODAP can handle any number of datasets within a single document and, by using "lookup" formulas, can relate one to another, more should be done to make it easier for learners to understand and make use of these facilities.

Journaling—In practice data scientists must document their work so that others can track all of what has been done with the data. A good software tool will help with this process by creating a human-readable and annotatable journal. Under what conditions should students be expected to create such documentation?

REFLECTIONS

Goals for Data Science Education at the School Level

I think what's important for school level students with regard to data is the experience of working with data much more than mastery of skills and concepts. Though it may not be the usual way of thinking about goals, my goal is for students to have regular, in depth experiences using data to undertake investigations or solve problems. I want students to emerge from these experiences convinced that data are fun, interesting, and useful. If they have become excited by what they have seen in data, and are convinced that the substantial work involved in getting data into shape is well worth the effort, then they will be likely to want to reach for data subsequently. Data habits of mind are probably more important than mastery at the school level because they constitute a way of thinking about the world in which data are central.

So this all leaves me skeptical about an approach whose goals revolve around introduction of a new discipline with a framework and a checklist of objectives. I would center curriculum development on wonderful datasets and activities that have students exploring, visualizing, model building, and searching for patterns. I would leave room for projects in which students have considerable choice and control. Imagine students who emerge from secondary school having had one or more intensive, data-rich immersions during each of their school years. Ideally these experiences would be intentionally interdisciplinary with their foci varying across many subject areas.

Course Versus Integration

It surprised me in our seminar that we appeared to begin with an implicit assumption that our focus was on a data science course at the secondary level without discussion of other possible strategies for bringing data science education into schools. I think it is very important that there be an explicit discussion of long-term strategies. Apparently, prior experience with computer education prejudiced the group against beginning with an integrated approach, but the conditions

that brought about these prior difficulties should be examined to determine their relevance to the current endeavor.

Software

Rob Gould's (2018) account of using RStudio as the primary data analysis software in the IDS course proved a persuasive argument for seeking alternatives. A basic tension arises between the desire to have students learn to use professional data analysis tools (RStudio, Tableau, …) versus environments designed for learning (Fathom, CODAP, INZight, …). I remain convinced that software designed for learning, provided it has sufficient data science capabilities, can give learners more meaningful experiences working with data because its interface can be kept simple and tuned to the evolving needs of the learner.

Certainly research on how students approach data science tasks and how they come to understand data "moves" will help us better understand what software learning environments should be like. CODAP, while a worthy start, needs additional capabilities and improvements to be a serious contender. Fortunately, CODAP is in ongoing development as it partners with projects (currently about 12) that provide funding, classroom testing, and guidelines.

Here is a list of considerations for CODAP's future development as a data science education environment.

- Feature: Residual plots (as in Fathom)
- Feature: Journaling that helps users document their work. Could be a bit "smart" about recognizing data moves. Consider a journaling plugin.
- Automation: Allow user to script a sequence of data moves so that user can create new tools.
- Statistical Modeling: As plugins
- Enhanced simulation

REFERENCES

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review 65*(2), 167–189.

Finzer, W. (2014). Hierarchical data visualization as a tool for developing student understanding of variation of data generated in simulations. In *Proceedings of the Ninth International Conference on Teaching Statistics*. Voorburg: International Statistics Institute.

Erickson, T. (2017 July 18). More about data moves—and R. [Web log post]. A Best-Case Scenario. Retrieved from https://bestcase.wordpress.com/2017/07/18/more-about-data-moves-and-r/

Gould, R. (2018). The Mobilize introduction to data science course: An overview**.** In Biehler, R. et al. (eds.). *Paderborn Symposium on Data Science Education 2017: The Collected Extended Abstracts*. Paderborn: University of Paderborn.

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., Tangmunarunkit, H., Trusela, L., & Zanontian, L. (2016). Teaching data science to secondary students—the Mobilize introduction to data science curriculum. In *Promoting Understanding of Statistics About Society. Proceedings of the IASE Roundtable Conference*.

LINKS

- CODAP: http://codap.concord.org
- RStudio: https://www.rstudio.com

DATA SCIENCE IN COMPUTER SCIENCE CLASSROOMS: A UNITED STATES PERSPECTIVE

Joanna Goode
5277 University of Oregon
Eugene, OR 97403-5277
goodej@uoregon.edu

This extended abstract provides a written companion to the presentation at the symposium focused on how data science has been infused in computer science education in United States schools. Particular attention is given to accessible curriculum and inclusive pedagogy that invites and supports a broad range of secondary student learners in engaging with data science curriculum. Part of this approach involves infusing sociocultural and sociopolitical connections as organizing topics in data science education.

COMPUTER SCIENCE IN SCHOOLS

In United States schools, data science has become infused as a central "big idea" in computer science education, and its reliance on computing has supported its growth as a central area of study in computer science education coursework. First introduced by President Obama, the national "Computer Science for All" educational movement has helped spread computer science learning opportunities across schools in the United States. Yet, despite this growth, the data science and computer science disciplines both suffer from the same constellation of equity and access issues in schools, universities, and industry. Descriptions of the historical and contemporary social context of computing in secondary schools demonstrates how issues of access, inclusion, and a lack of culturally responsive and engaging curriculum have sustained educational inequities (Margolis, Estrella, et al. 2017).

BROADENING PARTICIPATION IN COMPUTING

To address this problem of underrepresentation in computer science, nine years ago the National Science Foundation invested in the development of two courses designed specifically to broaden participation in computing in high schools: Exploring Computer Science (Goode, Chapman, & Margolis 2012; Goode, Margolis, & Chapman 2014); and AP Computer Science Principles (Astrachan & Briggs 2012). Rather than a singular focus on a coding language as was typical in year-long courses, computer science was presented as foundational in both courses and organized around "big ideas" relevant to students' lives. In both courses, data science is presented as one of a handful of foundational topics that organize computer science student learning progressions. In addition to framing computing as a topic that includes and goes beyond programming, both courses also offered a new way of thinking about student participation through the use of "computational practices." These "computational practices" are key areas of focus for teachers to nurture in students alongside the development of content knowledge.

- Analyze effects of computing
- Design creative solutions and artifacts
- Apply abstractions and models
- Analyze computational work and the work of others
- Communicate computational thoughts and processes
- Collaborate with peers on computing activities

These practices are useful in emphasizing not only facts and knowledge about computing, but also the dispositions and professional practices associated with performing content-area activities.

United States national and state policy efforts have also reinforced this more foundational and inclusive approach to teaching computer science. Recent national standards and frameworks

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

developed by non-profit groups and teacher organizations followed the trend of these high school courses and also put forward data science as a key topic in primary and secondary CS Education. The last year has seen the emergence of the Computer Science Framework (2016) and a reworking of the Computer Science Teachers Association Computer Science Learning Standards (CSTA 2017), both of which included data science as a core conceptual area and computational practices as core tenets in their organization of learning pathways. The Framework describes their "Data and Analysis" unit as "Computing systems exist to process data. The amount of digital data generated in the world is rapidly expanding, so the need to process data effectively is increasingly important. Data is collected and stored so that it can be analyzed to better understand the world and make more accurate predictions." (CS K–12 Framework 2016, p 90). Further, the framework breaks down this topic into sub-areas of collection, storage, visualization and transformation, and inference and models.

SOCIOCULTURAL AND SOCIOPOLITICAL DIMENSIONS

To bolster the data science curriculum in Exploring Computer Science, the Mobilize project in Los Angeles was created to deepen the data science learning activities and infuse participatory sensing experiences in the instructional unit. As part of a school-university partnership in Los Angeles, the Mobilize project involved a multi-dimensional effort to infuse data science in Los Angeles high schools. An initial effort of this Mobilize project infused data curriculum into existing Exploring Computer Science courses as an advanced instructional unit. The curriculum was accompanied by a robust set of professional development experiences, in which teachers themselves collected and analyzed data, and discussed pedagogical strategies for engaging all students in data science learning. After this professional development series, teachers implemented this new curriculum, along with an inquiry-based instructional approach, in their secondary classrooms. Researchers conducted weekly classroom observations across three ECS classrooms implementing the participatory sensing data science learning materials. A technical support team also supported teachers' use of new data analysis tools in classrooms.

The findings from this study showed how a participatory sensing approach to teaching data science led to a high level of student interest and engagement. We found that taking students through deliberate decision-making points with collecting, merging, analyzing, representing, and presenting data allowed students to better understand how such choices are at the heart of the sociocultural and sociopolitical angles of data science (Ryoo, Margolis, et al. 2013). Rather than viewing data as an abstract and objective entity, students were asked to make meaning of phenomena in their own lives through data. We discovered that teachers who employed culturally-responsive teaching practices captivated diverse groups of learners with the power and awe of data science. For instance, two vignettes from ethnically diverse classrooms in Los Angeles highlight the importance of preparing teachers to use culturally-responsive computing pedagogy that supports this type of inclusive classroom learning around data and computing.

Vignette 1: Connecting to Students and Assumptions of Data

The teacher shared a list of websites she had visited over a 24-hour period, emphasizing how her online activity could be considered data. She asked the class what the list might illustrate about her. Students noted things such as "you like gadgets" because she visited gizmodo.com, and that she liked "chismé" (Spanish for "gossip") because she visited gawker.com, and that she enjoyed laughing because gawker postings were often funny and entertaining. The teacher then asked the class: "What doesn't it tell you?" Students listed: "your race, your birthday, your social security number." The teacher asked the students what race this list suggested she could be; one student said that she seemed "white" because she visited computer-related sites. Several students agreed, to which the teacher replied, "Why is it that we associate being white with knowing a lot about computers?" One student shared, "Most computer geniuses are white." Students nodded heads in agreement, and the teacher replied, "So what are we [all people of color] doing here?" Student said, "Not being white." The teacher stated, "So look around this room. Most of us are people of color…we're trying to break the stereotype."

This vignette shows how the teacher makes the subject accessible by connecting computing and data science ideas to students' personal interests, experiences, perspectives, and

everyday lives. The teacher contextualizes the data, in this case her own data, within the cultural practices of going online. Students are encouraged to draw on what they notice and see in the teacher's data—a prompt that allows all students, regardless of prior data science knowledge, to engage in the discussion. Further, by connecting this data inquiry into larger structure of equity and systems of power, the teacher infuses a sociopolitical discussion through this lesson. The teacher also uses this lesson to exhibit how diverse groups of students have different experiences, even when using the same technology, because their data is shared in different ways and different assumptions are made about them.

Vignette 2: Importance of Cultural Context and Place

A second vignette highlights how the cultural and place-based context that participatory sensing leads to can connect with large sociocultural and sociopolitical community issues. In the case of one classroom, students agreed that they would engage in a project exploring nutrition and snacking. When introducing an instructional activity in which students would collect their own snacking information (e.g., what they were eating, why they ate it, how much it cost, etc.), the teacher introduced the concept of "food deserts." The teacher asked if students remembered what used to be in the nearby empty lot. One student remembered it had been a little farm. The teacher replied, "Whenever I pass that, I get a pit in my stomach because there used to be a farm on that lot—where your family, neighbors, and people in your community used to grow fresh food. Would you say that the lack of fresh foods in your community is a problem?" Discussion then turned toward how often students went to fast food chains and how they couldn't find fresh produce in their neighborhood. The discussion helped frame students' efforts collecting and analyzing data on snacking habits using computing tools, and reinforced the importance of insider knowledge of community and context when engaging in a locally-based data science project.

CONCLUSION

As data science continues to find a home in computer science education in the United States, exploring how supportive curriculum and pedagogy can support robust learning remains a central issue. The inclusion of a participatory-sensing approach in the foundational Exploring Computer Science course offers a promising approach to teaching data science content, infusing computational practices, and presenting data as a living sociocultural and sociopolitical artifact. Future efforts in secondary data science education should continue to explore the curricular and pedagogical hooks that continue to actively and intentionally involve girls and other underrepresented groups into classroom learning spaces.

REFERENCES

Astrachan, O., & Briggs, A. (2012). The CS principles project. *ACM Inroads*, *3*(2), 38–42.

CSTA. (2017). *CSTA K–12 Computer Science Standards*. New York: Association for Computing Machinery. Retrieved from https://www.csteachers.org/page/standards

Goode, J., Chapman, G., & Margolis, J. (2012). Beyond curriculum: The exploring computer science program. *ACM Inroads*, *3*(2), 47–53.

Goode, J., Margolis, J., & Chapman, G. (2014, March). Curriculum is not enough: The educational theory and research foundation of the exploring computer science professional development model. In *Proceedings of the 45th ACM technical symposium on Computer science education*, 493–498. ACM.

K–12 Computer Science Framework. (2016). Retrieved from http://www.k12cs.org

Margolis, J., & Goode, J. (2016). Ten lessons for computer science for all. *ACM Inroads*, *7*(4), 52–56.

Margolis, J., Estrella, R., Goode, J., Holme, J. J., & Nao, K. (2017). *Stuck in the shallow end: Education, race, and computing* (*Revised)*. Boston: MIT Press.

Ryoo, J. J., Margolis, J., Lee, C. H., Sandoval, C. D., & Goode, J. (2013). Democratizing computer science knowledge: Transforming the face of computer science through public high school education. *Learning, Media and Technology*, *38*(2), 161–181.

THE MOBILIZE INTRODUCTION TO DATA SCIENCE COURSE: AN OVERVIEW

Robert Gould and the Mobilize Team[1]
UCLA Dept. of Statistics, Los Angeles, CA 90095-1554, USA
rgould@stat.ucla.edu

Funded in 2010 by the National Science Foundation, the Mobilize Project is a partnership between the Los Angeles Unified School District (LAUSD) and several units in the University of California, Los Angeles (UCLA), including the departments of Computer Science, Statistics, and the Graduate School of Education and Information Studies. Launched as a pilot program in 2014-2015, the Introduction to Data Science (IDS) course became the most important product of Mobilize. The course was piloted by ten teachers in 2014 and expanded the following year to 25 new LAUSD classrooms. This academic year (2017-18), 27 LAUSD teachers are teaching 44 sections of IDS, and 17 math teachers from six new school districts in Southern California are teaching 27 sections of the course. IDS is designed to develop computational thinking and statistical thinking in a problem-solving setting, and its core goals are to raise students' awareness of the role of data in our culture and in their everyday lives, to prepare them to think critically about data, and to teach them to analyze data of a variety of formats and types.

This report provides an overview of the design of IDS and of the curriculum itself. Evaluation of the project focused primarily on the development of students' statistical reasoning, but this report will also discuss challenges in teacher preparation and development of effective pedagogy, as well as administration and logistical challenges encountered during implementation.

THE ORIGINS OF IDS

The Mobilize Introduction to Data Science curriculum (IDS) is a year-long high school course teaching data analysis in a computer-intensive environment. The primary goals of IDS are to introduce students to the role that data play in their lives and society, and to teach introductory approaches for analyzing and working with complex, modern data. IDS is situated in the mathematics curriculum in California – its state of origin – and in the parlance of the California Department of Education, the course "validates the Algebra II" requirement. More simply put, this means that the course has introductory Algebra as a prerequisite, and that students who successfully complete the course have satisfied the requirement to take intermediate Algebra. Because intermediate Algebra is required for admission to the two public university systems in California, IDS serves as an alternative pathway to college that bypasses Algebra II, a course that has been a hurdle for groups that have historically been underrepresented in the sciences (Burdman 2015).

The origins of IDS are unique and may have resulted in some idiosyncrasies in the curriculum. IDS was created somewhat late in the life cycle of the Mobilize Project, a partnership funded by the National Science Foundation (NSF). The partnership was between Center X of the Graduate School of Education and Information Sciences (GSEIS) at the University of California, Los Angeles (UCLA), the Center for Embedded Network Systems (CENS) in the Computer Science Department at UCLA, the Statistics Department at UCLA, and the Los Angeles Unified School District (LAUSD), the nation's second-largest school district (with over 650,000 students).

The original leaders of Mobilize were Deborah Estrin, Professor of Computer Science and Director of CENS, and Jane Margolis, Senior Researcher in GSEIS. (The author, a statistician, took over as the lead principal investigator of the Mobilize Project in 2012 when Estrin left UCLA.) Jane Margolis collaborated with Joanna Goode, Associate Professor of Education at the University of Oregon in creating the Exploring Computer Science (ECS) program (http://www.exploringcs.org).

[1] Suyenn Machado, LeeAnn Trusela, James Molyneux, Terri Johnson, Amelia McNamara, Jeroen Ooms, Jane Margolis, Jody Priselac, Joanna Goode, Derrick Chao, Hongsuda Tangmunarunkit, Steve Nolen, Kapeel Sable, Shuhao Wu, Maria Olivares-Pasillas

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

The primary purpose of ECS is to teach the fundamentals of computer science in high schools, and to encourage underrepresented minorities and girls to enter CS in particular and science in general.

One component missing from ECS was one that dealt with data, and part of the mission of Mobilize was to correct this. Estrin and her lab at CENS had developed the framework of "participatory sensing", a data collection paradigm similar in many ways to "citizen science" (Burke, et. al 2006). In citizen science, lay people ("citizens") collect data and send this data to researchers. Participatory sensing democratizes this structure by creating communities who collect data, but who also own and share the data they collect. A key element of participatory sensing is the mobile phone, which the developers of participatory sensing described as "a special and unprecedented tool for engaging participants in sensing their local environment" (Goldman, et. al 2008). Mobilize was proposed to the NSF as an attempt to implement participatory sensing in an educational setting, and to explore ways to use it to enrich the ECS curriculum, as well as other Science, Technology, Engineering, and Math (STEM) curricula by engaging students in data collection and analysis.

For the first few years, Mobilize developed and implemented short modules designed to leverage participatory sensing to enhance learning in Algebra, Biology, and ECS classrooms. The implementation included close cooperation between high school teachers and UCLA researchers to develop the modules and to teach the pedagogy behind them via an intensive professional development program. A key design constraint for the development of these modules was that they be equity-oriented (all students must participate and benefit equally) and inquiry-based (built upon a constructivist, active-learning model). The units were between 3 and 6 weeks long.

There were many reasons that these units were not resoundingly successful, but a primary reason identified by the Mobilize team was that participatory sensing is time-consuming and difficult. Essentially, the curriculum must provide time for students to develop basic data analysis skills and the conceptual understandings that support these skills, while also making time to collect data and to learn how to use the participatory sensing data collection app. Teachers must invest time in setting up the technology, and in learning to use it to manage their classroom data. More time would allow for students to have the space needed to absorb some difficult concepts, and would make the teachers' investment in learning the technology more worthwhile.

IDS was designed to be a course that would build on the successes of the early Mobilize units while providing the time and pedagogical development needed for the Mobilize lessons to be effective. One hindrance to a longer course was the California State Mathematics standards, which were not very flexible, and which made it unlikely that a new course would succeed unless it focused solely on traditional mathematical skills. Fortunately, in 2010 California adopted the Common Core Standards in Mathematics (Common Core State Standards Initiative 2010) and implementation began in 2014, just when IDS was piloted in LAUSD high schools. The new standards had a greater emphasis on mathematical and scientific modeling, and a stronger emphasis on data analysis, which made a course like IDS an ideal fit. Another factor in its favor was that LAUSD was able to get state approval for IDS to validate the Algebra II requirement. This meant that IDS could serve as a pathway to college and thereby guarantee a larger audience

IDS was first introduced to ten LAUSD teachers and implemented in their classrooms in 2014-15. The following year, six of those teachers became "teacher leaders" who worked closely with the IDS developers to revise the curriculum and to prepare a new cohort of 25 teachers in 2016. As of the 2017-2018 academic year, the course is taught by 44 mathematics teachers in Southern California (27 LAUSD teachers, 17 teachers from other districts). LAUSD serves a primarily economically disadvantaged population, with roughly 84% of students in the district living below poverty level. Ninety-four languages are spoken by LAUSD students, approximately one-fifth of whom are not fluent in English. The racial composition of the initial cohort of IDS students was reflective of the district population, though somewhat less diverse, having a lower proportion of Whites and African Americans. Approximately 90% of IDS students were Hispanic, 4% African American, and the remaining 6% were roughly equal proportions of White, Asian, Pacific Islander, and Native American.

THE IDS CURRICULUM

Development of IDS was informed by the American Statistical Association's Guidelines for Assessment and Instruction in Statistics Education K-12 Report (GAISE K-12) (Franklin et. al 2007). In particular, IDS emphasizes the beginning and intermediate levels (levels A and B) in that report, with the intention of preparing students for a future Statistics course that would emphasize level C (e.g., the AP Statistics course).

The IDS curriculum is organized into four units, each lasting approximately nine weeks. The units focus on broad themes. Unit 1 focuses on data. Students learn how data are stored, structures for storing data, and descriptive statistics. Unit 2 teaches informal inference (Makar & Rubin 2009) using randomization-based testing and introducing fundamental notions of probability. The third unit is about data collection. In addition to the usual suspects - observational studies, controlled experiments, random assignment, random sampling – Unit 3 also explores the advantages and disadvantages of participatory sensing, finding and "scraping" data from the internet, and other nontraditional data collection techniques. Unit 4, the final unit, focuses on multivariate modeling, with an emphasis on making and evaluating statistical predictions.

Daily class meetings have two flavors. On most days, students engage in active inquiry-based lessons to learn methods and develop conceptual understanding. At least once per week, students meet in a computer lab in which they use the R language (R Core Team 2017) to analyze data.

Software, Technology, and Big Data

The Mobilize project iterated through several statistical analysis software platforms before choosing R for the primary tool in IDS. R is implemented via RStudio (RStudio Team 2015) and is run on a central server, which allows the project to maintain a uniform classroom implementation and to troubleshoot problems quickly. The primary appeal of R was that it provides opportunities for students to write basic code and to thereby experience some fundamentals of software programming. A special package, mobilizR, was developed to allow students to code with a more unified syntax than is typically provided by R (https://github.com/mobilizingcs/mobilizr). This package includes "wrapper" functions that simplify commonly implemented multi-step commands into a single command. The mobilizR package was based on mosaic (Pruim, Kaplan & Horton 2015), a package designed to help undergraduate students implement R in statistics, math, and science courses.

The curriculum also makes use of the Mobilize “dashboard”, a tool for visualizing the multivariate data collected through participatory sensing. (See https://sandbox.introdatascience.org/). The dashboard is just one part of a suite of technology developed by Mobilize to implement participatory sensing in a classroom setting (Tangmunarunkit et. al 2015). The suite also includes software that allows students to manage their own data (including toggling privacy settings), and for teachers, in turn, to manage student data. The suite maintains data security so that data collected in one class cannot be viewed in another.

Figure 1 shows a snapshot of one component of the dashboard called the "snack board." The data in this particular visualization were collected by students whenever they ate a snack during a chosen week. (The data combine several classes and also include data from a few teachers during professional development sessions. Data have been anonymized.) In this case, the dashboard has been set to highlight cereals eaten at night. The dashboard is displaying photos of the cereal, the distribution of cost, the location in which it was eaten, wordclouds about why the snack was eaten and what was eaten, and the perceived health level of the snack. Although this visualization is deliberately busy and crowded, each component can be toggled on and off. Further, the dashboard is interactive. Clicking on any component updates all other displays to focus only on observations sharing the selected values. Thus, after clicking on the correct sequence of displayed values, students could choose to focus on cereals eaten at night that cost less than one dollar, and which had a perceived health level of 2.

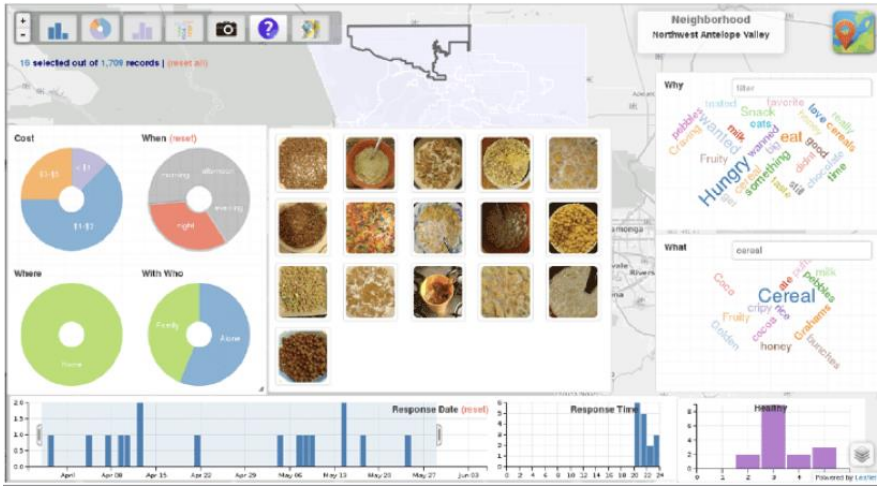


Figure 1. Dashboard visualization of snacks

The data collected by a class engaged in participatory sensing is typically not large in terms of the number of independent observations. However, the data share many characteristics of data that might be classified as falling under the Big Data tent. The participatory sensing data are of multiple types (words, categories, numbers, photos, GPS coordinates), they are complexly structured (individual students providing multiple observations across time and space), and they are collected via a non-traditional sampling paradigm that doesn't allow for traditional inference approaches. The complexity, structure, small size, and lack of a random sampling scheme provide challenges for the classroom. Despite these challenges, these data can provide opportunities to engage students in notions about data privacy and security, the complexities of inference, and the challenges of measuring complex phenomena.

As promising as participatory sensing is, it is not sufficient to support an entire data analysis curriculum, and so IDS is supplemented with data gathered from a variety of sources, including the Youth Behavior Risk Survey (https://www.cdc.gov/healthyyouth/data/yrbs/data.htm) the American Time-Use Survey, (https://www.bls.gov/tus/), data scraped from various movie-related websites, and data collected from sports websites. Many of these datasets are included within the mobilizR package. Whenever possible, these datasets are explicitly tied to participatory sensing campaigns. For example, after spending time collecting data about their own use of time using the participatory sensing paradigm, students are introduced to data from the American Time-Use Survey, which they analyze.

Statistical Modeling with the Data Cycle

A central component of the curriculum is the "Data Cycle" (Figure 2), which is a graphical representation of the four steps of the statistical investigation process as defined by GAISE (Franklin et al 2007). The statistical investigation process is itself a simplification of the PPDAC cycle of Wild & Pfannkuch (1999). The Data Cycle modifies these four steps slightly by replacing the "collect data" step with the more generic and data science-relevant "consider data" step.

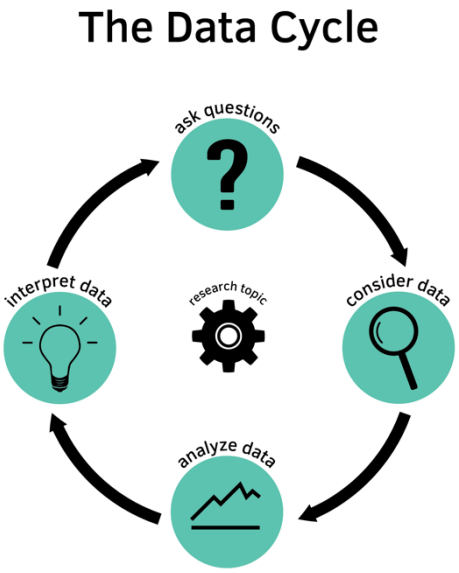


Figure 2. The Data Cycle

The Data Cycle was introduced to correct issues identified through student work from the original Mobilize modules. The Data Cycle is intended to serve as a reminder to both teachers and students that a statistical investigation consists of more than simply typing the correct code and printing out a graph. It reminds students that an analysis serves a purpose: to answer questions. Posing productive statistical questions turned out to be a challenge for both teachers and students. We found that many of the mathematics teachers we worked with were unfamiliar with the notion of posing questions that could be answered with data (i.e., "statistical investigation questions", using Arnold's (2013) terminology). In one case study, we saw evidence that suggests that teachers' failure to phrase productive questions early in an investigation may be detrimental to the investigation (Gould, Bargagliotti, Johnson 2017). Therefore, a significant amount of professional development time is devoted to this topic.

Topics in the Curriculum

The topics covered in the curriculum include those that are quite common in many secondary level statistics courses, and some that are typically offered only in more advanced statistics courses. For example, in IDS students learn to compute probabilities using the normal distribution, although they use R as their calculator. As do many other statistics students, they learn how to make and interpret histograms, but they also study the algorithm that generates histograms.

IDS covers more advanced topics that are not common in traditional introductory statistics, such as data privacy issues, data classification through k-means, and prediction using Classification and Regression Trees. One lesson that addresses data privacy asks students to keep a "data diary" for 24 hours, recording every activity they perform which contributes to their "data trail". This activity follows a video about data privacy violations and a classroom discussion about how and where data are recorded about our activities.

CHALLENGES

Implementation of IDS was, and continues to be, challenging for a number of reasons. Two prominent issues are a need for specialized professional development and the lack of validated assessments for learning/understanding in the realm of data science.

Teacher Preparation

IDS requires substantial professional development resources. In California, IDS is classified as a mathematics course, even though it predominately features topics in computer science and data analysis. This classification means that teachers must have a mathematics credential, and many mathematics teachers are not well prepared to teach data analysis.

Understandably, few teachers in any discipline are familiar with R, and R has a famously steep learning curve. For these reasons, professional development is required and a substantial portion of professional development time must be spent on learning the technology, in addition to course content and pedagogy.

IDS also presents teachers with some pedagogical challenges. Many secondary school mathematics teachers are not comfortable with situations in which the answers to questions might vary depending on the data collected, or with problems for which there might be several plausible and differing interpretations of the same analyses. Analyzing data about real life can raise issues that are rarely raised in mathematics classes, and sometimes lead to off-the-cuff statements that can insult or over-generalize. (Two examples that I personally overheard during classroom visits while students were discussing their findings: "Everyone knows Mexicans are short" and "Only white people buy bottled water.") Teachers need to be prepared for potentially difficult discussions and need to be shown how to engage students without any single student being excluded or diminished (either by the teacher or by other students). For these reasons, professional development includes strategies for "data discussions" to produce thoughtful, collaborative discussion.

Research on Data Science Education

Data science is a new and emerging field, and it is fair to say that there is no consensus definition of "data science". We designed our course as a blend of data analysis with a heavy dose of computational thinking. While there exist literatures on statistics education and on computer science education, the former emphasizes formal statistical inference, and there is little research about the intersection of statistical thinking and the computer. There do exist validated assessments of statistical thinking that include data analysis (e.g., the LOCUS (Jacobbe, Case, and Whitaker & Foti 2014), but these do not include technology. There are assessments of computational thinking (https://www.sri.com/work/projects/principled-assessment-computational-thinking-pact), but these do not include applying computational thinking to data analysis.

For our own purposes, we used the LOCUS (Levels of Conceptual Understandings in Statistics) as a pre- and post- tool to measure change in statistical thinking. The LOCUS is tied explicitly to the GAISE levels of beginning, intermediate, and advanced, and measures across four areas that map (although somewhat imperfectly) to the Data Cycle: Asking questions, collecting data, analyzing data, and interpretation analyses. An analysis by the project's external evaluators indicated statistically significant gains in the total LOCUS score, although these varied from teacher to teacher. Interestingly, the greatest gains appeared to be from students who scored high on the pretest, suggesting that some students had more experience with statistics, and that this experience was helpful in learning data science.

CONCLUSION

While interest in IDS is growing, this course (and any course teaching data science) faces challenges. For example, we continue to be challenged by how to teach R while simultaneously teaching statistical thinking. While the "regular" classroom instruction is often lively and interactive, teachers report that the computer lab tends to be quite subdued, with students simply trying to answer questions without much concern for explaining their reasoning or grappling with problem-solving. By some reports, many students in the computer labs turn to a student "expert" (and it seems every class has one or two) who "gets it", and they simply copy-and-paste what that expert has done. We are slowly improving this process; currently we are piloting a pair-programming (McDowell, Werner, Bullock, and Fernald 2002) format for the computer labs, hoping that a structure for collaboration will not only foster more thoughtful work on the part of the students, but will also allow teachers to identify areas where students struggle, and to acknowledge student success.

Another approach is to teach using a simpler, more accessible statistical analysis package and then, later, when students have assimilated some statistical concepts, introduce R (or Python, another software used by data scientists.) A challenge in using any software in a large, urban school district such as LAUSD is that there are few resources for technical support. Software must be inexpensive (R and RStudio are free) and must run on a central server so that software does not

need to be downloaded to and maintained on individual machines. The Common Online Data Analysis Platform (CODAP) shows some promise in this regard (Concord Consortium 2018). Currently, it supports hierarchically structured data and a variety of multivariate visualization techniques, including mapping. As development continues, CODAP could play a very important role in data science education. Looking slightly farther into the future, there is a need for a software pathway that, in the words of McNamara (2017), "bridges the gap" between data science learners and data science professionals.

Still, the greatest challenge is professional development. Professional development is expensive and, at least in California, schools are increasingly reluctant to pay for it. These costs could be mitigated if programs that prepare teachers paid more attention to both data analysis and programming.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Number 0962919. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

IDS is extremely grateful for its teacher leaders who have shaped the curriculum into what it is, and who continue to prepare future data science teachers: Pamela Amaya, Monica Casillas, Heidi Estevez, Joy Min, Robert Montgomery, Roberta Ross, Carol Sailor, and Alma Villegas-Torres. We are also, of course, extremely grateful for former principal investigators who got the project off the ground: Deborah Estrin (Computer Science), Mark Hansen (Statistics), Jane Margolis (Education), Jody Priselac (Education), and Todd Ullah (LAUSD).

REFERENCES

Arnold, P. (2013). Statistical investigative questions: An enquiry into posing and answering investigative questions from existing data (Doctoral dissertation). Retrieved from University of Auckland Research Repository - ResearchSpace. (Identifier: http://hdl.handle.net/2292/21305)

Burdman, P. (2015). Degrees of freedom: Diversifying math requirements for college readiness and graduation (Report 1 of a 3-part series). Oakland, CA: LearningWorks and Policy Analysis for California Education, PACE. Retrieved from Institute of Education Sciences ERIC collection. (ERIC Number: ED564291).

Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M.B., (2006). Participatory Sensing. *WSW'06 at SenSys*., Boulder, CO.

Common Core State Standards Initiative (2010). Common core state standards for mathematics. Washington, DC: National Governors Association Center for Best Practices & Council of Chief State School Officers. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Concord Consortium (2018), CODAP software. https://concord.org.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). Guidelines for assessment and instruction in statistics education (GAISE) report. Alexandria, VA: American Statistical Association.

Goldman, J., Shilton, K., Burke, J., Estrin, D., Hansen, M., Ramanathan, N., Reddy, S., Samanta, V., Srivastiva, M. (2008). Participatory Sensing: A citizen-powered approach to illuminating the patterns that shape our world. Retrieved from http://escholarship.org/uc/item/19h777qd

Gould, R., Bargagliotti, A., Johnson T (2017). An Analysis of Secondary Teachers' Reasoning with Participatory Sensing Data. Statistics Education Research Journal. 16(2). Retrieved from https://iase-web.org/documents/SERJ/SERJ16(2)_Gould.pdf

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014). Establishing the content validity of the LOCUS assessments through evidence centered design In K. Makar & R. Gould (Eds.) Proceedings of the 9th International Conference on Teaching Statistics.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. Statistics Education Research Journal, 8(1), 82-105. Retrieved from https://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1).pdf#page=85

McDowell, C., Werner, L, Bullock, H., and Fernald, J. (2002). The effects of pair-programming on performance in introductory programming course. *Sigcse '02 Proceedings of the 33rd SIGCSE technical symposium on Computer science education*, (pp. 38-42), Cincinnati, KY. https://doi.org/10.1145/563517.563353

McNamara, A. (2017). Considering the Gap Between Learning and Doing Statistics in *International Handbook of Research in Statistics Education*, Ben-Zvi, D., Garfield, J., & Makar, K., eds. Springer International Handbooks of Education, pp 463-465.

Pruim, R., Kaplan, D., Horton, N., Creativity, M., & Minimal, R. (2015). Mosaic: Project MOSAIC statistics and mathematics teaching utilities. R package version 0.10.0. [Computer Software.]. Retrieved from https://cran.r-project.org/web/packages/mosaic/index.html

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc. [computer software]. Boston, MA. Retrieved from http://www.rstudio.com/

Tangmunarunkit, H., Hsieh, C.K., Longstaff, B., Nolen, S., Jenkins, J., Ketcham, C., Selsky, J., Alquaddoomi, F., George, D., Kang, J. & Khalapyan, Z. (2015). Ohmage: A general and extensible end-to-end participatory sensing platform. ACM Transactions on Intelligent Systems and Technology (TIST), 6(3), 38.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. International Statistical Review, 67(3), 223-248.

"BIG DATA" AND STEM LITERACY THROUGH INFOGRAPHICS

Andee Rubin
TERC, Cambridge, MA, USA
Andee_rubin@terc.edu

This contribution has two parts: 1) some thoughts on how to choose appropriate data sets to introduce students to data science and 2) a brief description and lessons learned from a United States high school project: STEM Literacy through Infographics

THE THREE BEARS OF DATA

There is much talk about "big data" and how it will change many aspects of our lives, from the workplace to the shopping mall to the doctor's office. Big data are often characterized with the 4 V's: Volume (there are a lot of data points); Velocity (the data, especially when generated automatically, arrive quickly); Variety (data may include text, numbers, images, video, sound); and Veracity (data quality is sometimes marginal; a small percentage of incorrect data is expected). When we think about data science education, we sometimes assume that students should encounter big data sets immediately, in order to get an authentic "data science" experience. However, truly large data sets can be overwhelming to novices, and they are likely to have trouble seeing the forest for the trees, as they get distracted by many complexities that arise in cleaning, organizing and otherwise wrangling the data.

On the other hand, the data students generally encounter in school could be considered "small data." Often there are just a few questions, each of which has only a few possible answers (e.g., what's your favorite color?). Even if the individual questions are relatively open-ended and interesting, there are frequently no meaty relationships among the variables, so analysis begins and ends with looking at univariate distributions. Interacting with these kinds of datasets seldom gives students a glimpse of the power and excitement that data analysis can yield.

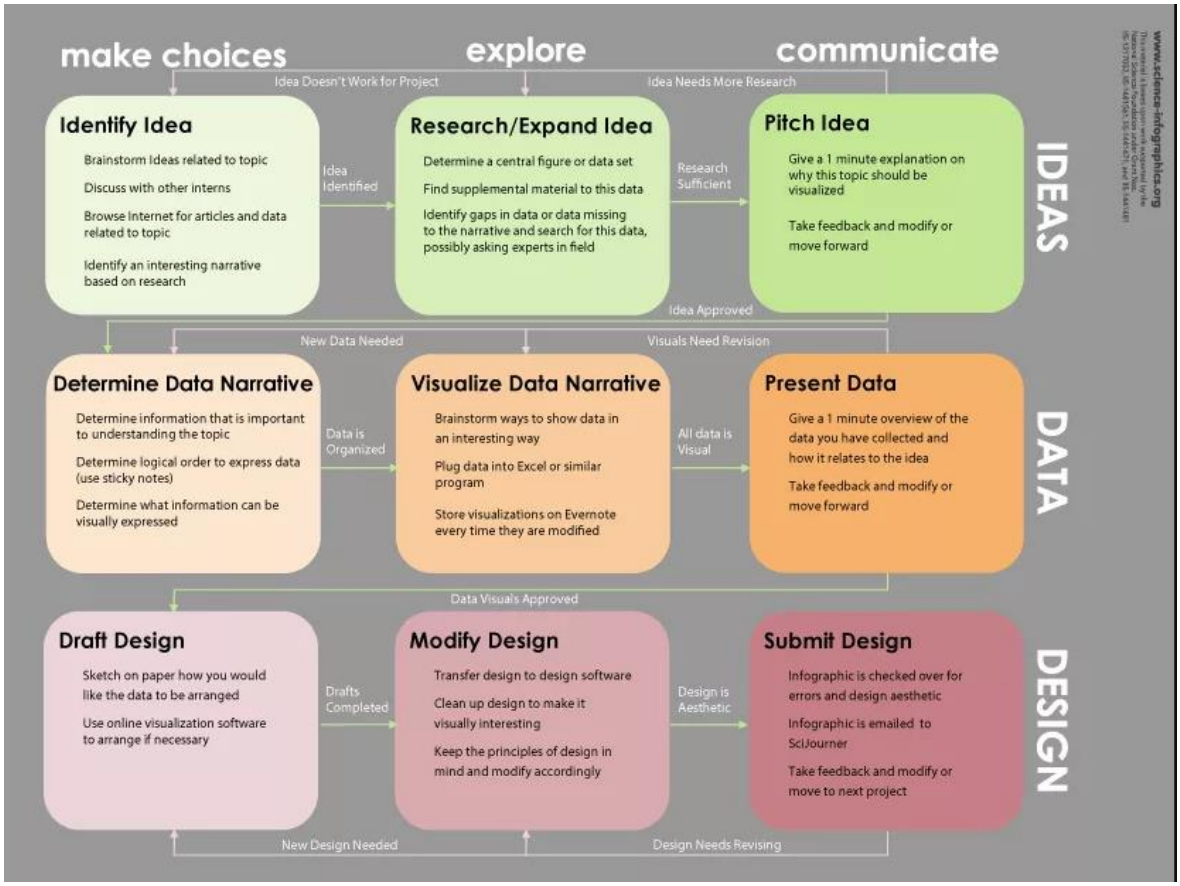
We find ourselves in a situation analogous to Goldilocks in the Three Bears fairy tale: one chair is too big and another is too small. Goldilocks eventually finds a "just-right" chair, and we should be searching for criteria for "just-right" datasets. As a start, it seems clear that such datasets would have to be multivariate, contain multiple types of variables and have within them rich relationships among variables, some of which might be true only for subsets of the data, allowing for different students to investigate different possible patterns. An international community can help vet and share such datasets as we find and prepare them for students.

STEM LITERACY THROUGH INFOGRAPHICS

STEM Literacy Through Infographics (SLI) is the third in a series of United States National Science Foundation-funded projects whose aim was to support high school students in seeing the relevance of science through journalism practices. The first two projects involved students primarily in text journalism, but the latest project integrates a focus on visual representations of data, as students create infographics (broadly, visual images such as charts and diagrams used to represent information or data) to communicate about a variety of scientific topics. The SLI project website (science-infographics.org) contains all of the resources described here, plus many more, which would be useful to anyone designing a data science course. It also lists the research papers the SLI project has written and contains links to many of them.

The SLI project conceptualizes students' creation of infographics in three broad categories dealing with Ideas, Data, and Design. In each of these categories, there are three phases: making choices, exploring, and communicating. (See Figure below.) While any individual student's infographic-creation process is likely to move between categories and phases, it has been helpful to students to provide some structure to the complex task of designing and implementing a product over several days or even weeks. The SLI project provided a unique venue for students to publish their infographics—an online website called SciJourner (http://www.scijourner.org), which contains both traditional text articles and infographics on a wide variety of topics. In order to publish on SciJourner, though, students had to revise their work until it met journalistic standards set by the SciJourner editor (who had actually been a scientific journal editor earlier in his life).

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374



The SLI project did not produce specific curriculum; rather, we developed a set of curricular resources and support materials, made recommendations about appropriate technology for students to use, and provided professional development for participating teachers, who taught subjects ranging from ecology to language arts. Then we tracked and analyzed what happened in these teachers' classrooms. In the three years of the SLI project, the team noticed where students struggled and created additional resources to meet their needs. Here we describe two examples of such resources that are particularly relevant to data science education.

Restructuring data

We recognized that students needed two different digital tools to create their infographics—one for exploring data and one for graphic design, which we called a "canvas tool." Most of the teachers in the project chose to use Google Sheets as the tool for exploring data (in spite of our suggestion that they use CODAP) and Venngage, a free online infographic creation tool, as their canvas tool. Students generally read data into Google Sheets from the Web, then tried to figure out how to make the particular graph they wanted, which often required changing the structure of the data to match Venngage's requirements. For example, students might import data from Gapminder, in which each case is a country and the years are variables, and want to make a graph comparing two countries over time. To create this kind of graph in Venngage, the cases have to be years and the countries are variables. Realizing that creating this graph requires a change in the structure of the data and knowing that "transpose" is a function that will accomplish this change was beyond most students (and teachers). In response, we developed a set of "data manipulation tutorials" describing situations such as these and explaining how to use Google Sheets to restructure data.

Data visualization pitfalls

Students who participated in SLI varied greatly in their experience with data and graphs, but many of them had few previous opportunities to create their own data visualizations. At the same time, Venngage offers a wide variety of graphs, including some that are quite esoteric and prone to misuse. This combination often led to students' creating inappropriate graphs and/or making common representational errors. For example, one group of students made a stacked bar graph that included in a single bar the average number of hours high school students in four different grades slept. (See Figure below.) Their goal was to compare the four values to make a statement like, "12th graders sleep less than other high school students," but the easy availability of stacked bar graphs lured them into making a graph that made little sense. Students also often used pie graphs inappropriately, combining frequencies that did not add up to 100% or did not represent mutually exclusive categories. In response to these observed issues, we are creating a series of short documents, each of which provides examples of "pitfalls" associated with a common graph type. The intent is for these documents to be resources that a teacher can use, providing them to students whom s/he sees has made an error creating a particular graph.

CONCLUSION

All of these resources are and will be available on the SLI website for the foreseeable future. The project team hopes that this legacy will be useful to teachers, students, curriculum designers and researchers after the project ends.

REFERENCES

Enyedy, N., Polman, J. L., Graville Smith, C., Bang, M., Warren, B., Rosebery, A. S., Burke, J., Wagmister, F., Bolling, A., Fitz-Gibbon, T., Halverson, E. R., & Nasir, N. (2014). Learning and becoming through art-making: Relationships among tools, phenomena, people, and communities in shaping youth identity development. In Polman, J. L., Kyza, E. A., O'Neill, D. K., Tabak, I., Penuel, W. R., Jurow, A. S., O'Connor, K., Lee, T., & D'Amico, L., (eds.) *Learning and becoming in practice: The International Conference of the Learning Sciences (ICLS) 2014, Volume 3* (pp. 1436–1445). Boulder, CO: International Society of the Learning Sciences.

Polman, J. L., & Gebre, E. H. (2015). Towards critical appraisal of infographics as scientific inscriptions. *Journal of Research in Science Teaching*. doi: 10.1002/tea.21225

DATA SCIENCE AND DATA LITERACY IN SCHOOL: OPPORTUNITIES AND CHALLENGES

Sue Sentance
King's College London
sue.sentance@kcl.ac.uk

With data science becoming increasingly topical, there is a need for more young people to be given access to the relevant knowledge and skills to become practitioners in this area. Introducing data science education into school could have a positive impact on the pipeline problem which industry faces, as well as giving all children key data literacy skills to support any future life journeys. Data science is an interdisciplinary field that incorporates aspects of statistics and computer science, and therefore could be taught in school within either mathematics or informatics (computing) classes (or both). Data literacy can be seen as an important forerunner to data science, and incorporates key skills in identifying, collecting, and analyzing data. In this extended abstract, the situation of school data science from the perspective of the computing curriculum is discussed, with particular regard to issues of content, structure, challenges, and teacher education. The computing curriculum in England is used as exemplification, but a rather disappointing lack of focus on data literacy in this curriculum means that some of the ideas presented here are somewhat speculative.

INTRODUCTION

All students learn mathematics in school and in some countries students also study computing (or informatics, computer science, ICT—variations of the subject have a range of different names). In terms of this discussion, both areas of the curriculum are important for learning data science. Here I am focusing only on the computing in the curriculum and its contribution to an education in data science.

Data science is a broad and transdisciplinary field incorporating statistics and computer science and many other disciplines including sociology, communication, and management (Cao 2017). Originally proposed within the statistics and mathematics community where the focus was primarily on data analysis, data science today includes data mining, machine learning, and more, and has become an intrinsically complex field (Cao 2017). A forerunner to data science could be said to be data literacy, which means the ability to identify, collect, evaluate, analyze, interpret, present, and protect data (ODI 2015).

The Royal Society, a prestigious body in the UK, recently published a report on machine learning (The Royal Society 2017a) where it highlighted the importance of machine learning and data science in our future lives. Within this report was a recommendation that appropriate skills to understand this area be introduced into the school curriculum from primary level to age 18. An even more recent, and potentially highly influential, report on Computing education by the Royal Society (The Royal Society 2017b) reiterated this recommendation. The fact that a prestigious body such as The Royal Society is calling for machine learning, and by association data science, to be incorporated into the curriculum, suggests that this should be a consideration for computing curriculum developers. In this extended abstract I discuss one current Computing curriculum and how it might need to evolve to provide the foundations of data science education for students.

THE SCHOOL CURRICULUM IN COMPUTING

In England a curriculum for Computing was introduced in September 2014, replacing ICT with which it had some overlap. The new elements are quite substantial and relate to computer science including computer programming, networks, computer architecture, and data representation. The introduction of Computing for all children from the age of 5 is important to note as it demonstrates that the subject is needed by *all* children and not just a self-selecting few. Further along the school timeline, there are qualifications in computer science that are optional including a *GCSE* which can be taken at age 16 and an *A-Level* which can be taken at age 18. In this section I will discuss what data science and/or literacy means in a mandatory school curriculum (for everybody), and later address assessment issues relating to qualifications.

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

Computing is an obvious area of the curriculum in which to develop key skills that could provide building blocks for the learning of data science. These building blocks are:

1. Data collection and analysis
2. Algorithms and programming
3. Ethics and moral issues

There are of course other aspects of knowledge and skills which will be taught within computing but the three given are the topics that will support further study in data science. Let us consider each of these school topics in turn.

Data collection and analysis

Under this heading, students could learn how to identify and collect data, ensure that it is ready for analysis and then carry out simple analysis of the data. Understanding data is part of computational thinking (Barr & Stevenson 2011) which makes it a fundamental part of the computing curriculum. Many activities can be incorporated into primary and secondary computing lessons where students consider the type of data they are collecting, how it can be stored in a computer, and use software applications to analyse it and generate information.

Algorithms and Programming

Under this heading, students learn how to use a programming language and how to develop algorithms that can be automated through the use of a programming language. In the curriculum in England, students learn by the age of 14 to "*design, use and evaluate computational abstractions that model the state and behaviour of real-world problems*" and to "*understand several key algorithms that reflect computational thinking*" (DfE 2013). They also need to write programs in two different languages and all of these skills are extended in the elective courses from 14–16 and from age 16–18.

Ethics and moral issues

Under this heading, students consider the impact of automation on society and individuals, and how the way that algorithms can be used make predictions or recommend decisions may impact people profoundly. Students also need to understand the impact that collecting large amounts of data about people has on privacy.

With these building blocks in place, students could be well placed to undertake more specialized study of data science later on in their school careers. In the next section two projects are described that illustrate how these topic areas can be introduced to young people in exciting ways.

EXAMPLES OF ACTIVITIES INVOLVING DATA IN SCHOOL

There are some existing initiatives being carried out in the area of data and education. Two examples are given here.

Our Data Ourselves

A project at King's College London, called **"***Our Data Ourselves***"**[1] has been designed to broadly consider the personal data generated in the everyday lives of young people through their mediated, cultural, and communicative practices. The project aims to contribute to a deeper understanding of our information-rich environment and the unprecedented growth of the data that is personally generated. The project engages with young people between the ages of 14 and 18 to develop tools and applications to visualize components of the Big Social Data (BSD) that they generate (Pybus, Coté & Blanke 2015). The project team have run workshops where young people reverse engineer apps to understand what happens to the data they generate through using their mobile phones.

[1] https://big-social-data.net/about/

Urban Data School

Another project exemplifying exciting work with data is the Urban Data School[2]. This project involves working with schools to enable pupils to critically analyse complex Smart City data sets, and to use data literacy skills to investigate critical urban issues. The work of the Urban Data School covers several aspects: use of data sets, developing data skills, enabling pupils to tell data stories, and promoting innovation. Trials have been carried out in primary schools where children interpreted data visualizations related to energy consumption and generation through solar PV (Wolff et al 2016). This is an exciting initiative that is part of the wider Smart Cities programme across the UK.

Both these initiatives are forging the way in enabling us to see how data literacy, a cornerstone of data science, can be brought to young people, but there are challenges in bringing activities such as these into a mandatory curriculum.

CURRENT CHALLENGES AROUND TEACHING COMPUTING IN SCHOOL

I have argued that the provision of computing in the curriculum is important in the discussion around data science education in school. However, for countries where computing does not yet exist as a discrete subject in the school curriculum, it is important to highlight the challenges that introducing computing as a subject for all students brings. Three main challenges that can be identified from the experience in England relate to teachers, the balance/breadth of the curriculum, and gender balance.

Firstly, **teachers** need to be confident in the subject matter being taught and how to teach it. In England the teaching of the algorithms and programming areas of the curriculum has caused teachers the most difficulty in recent years. These areas are important, but teachers need plenty of time and training to get up to speed. Training teachers is expensive and quite often doesn't happen. Teachers without enough training lack confidence. New teachers are hard to find and many do not stay long in computing; the recruitment problem in Computing has been highlighted in the recent Royal Society report (The Royal Society 2017b).

Secondly, it is important to ensure that computing instruction is **balanced** around all areas of the curriculum; it is probable that this balance has not been achieved in the context of the curriculum in England. Programming is found to be the most difficult aspect of computing and there is perhaps a limited understanding by policymakers and examination boards as to what can reasonably be achieved at what level. It is important to balance the study of algorithms and programming with the other building blocks described above: data and ethics provide for a balanced curriculum that would facilitate the development of broad skills across computer science, information technology, and digital literacy. In England this is the desire but not the reality: future revisions of the curriculum could address this.

Thirdly, in the UK and elsewhere there is a lack of **gender** balance with very few girls choosing computer science (rather than computing) once the subject becomes elective. There are many reasons for this that have deep societal roots; however, it may be that an increased focus on working with large datasets and using students to ask key social questions may increase motivation and interest across the genders for computing. Currently in the UK only 1 in 5 computer science students age 14–16 are girls and 1 in 10 from 16–18 (The Royal Society 2017b), so there is huge scope for improvement.

TEACHERS AND DATA SCIENCE

Teachers' perspectives

We carried out a small-scale survey of teachers (n = 36) to canvass opinions on data science in the Computing curriculum in the UK and this revealed that, although teachers feel it is an important topic, there is concern that an already crowded curriculum has no room for any additional material. In addition, pressure on teachers from frequent recent changes may mean there

[2] https://www.urbandataschool.com/

is limited appetite for more change. This highlights the need for a rationalization of what we can and cannot include, and the need to carefully prepare for the implications for existing teachers.

Teachers gave a variety of reasons for not delivering data science in school. The most common sentiment was that there was already too much in the curriculum to add any more content: "*It's a crowded curriculum, what would we lose? We already lack databases at KS3 and KS4, they would be far more foundational to student learning."* This statement refers to the fact that the move from ICT to Computing took some emphasis away from databases, which had previously been included, and moved the focus to programming. Another teacher commented that "*It is one of many interesting topics that we should AVOID overloading the GCSE curriculum with."* On the other hand, some teachers reported that the subject was interesting but that it would need to have a qualification developed in it and resources made available.

In answering other questions, teachers felt that the implications of big data, machine learning, and artificial intelligence were important topics that should be incorporated into the computing curriculum. Computing teachers were asked about including data science in the mathematics curriculum, but no comments on this were received (positive or negative).

Implications for teacher education

Anything new in the curriculum will have an impact on teacher education. It is a salient fact that no teacher will have learned data science in school. Teacher education, supply, and retention are already issues in computing education. To facilitate data science education, resources are needed to support teachers and teachers will need training. Content needs to be accessible to teachers who may not yet be familiar with the techniques of data science. Work would be needed to establish links between mathematics and computer science teachers; this could be potentially very exciting, but possibly difficult to implement.

ASSESSMENT AND DATA SCIENCE EDUCATION

In a country where there are already qualifications in Maths, Statistics, and Computer at age 16, the question is where data science, if included in the curriculum, would fit into the assessment process. In England it would be difficult to justify teaching a subject area which did not have accompanying summative assessment. Here we consider various optional ways of incorporating data science as an examinable subject for students aged 14–16 (grades 9–10):

Option 1.	The creation of a new subject/qualification called Data Science with aligned examinations at age 16.
Option 2.	Incorporating the teaching of data science within the computer science course, but assessing via a separate qualification at age 16 (this is the way that Statistics is assessed, as it is taught within Mathematics).
Option 3.	Modify the computer science course to incorporate data science aspects and remove some elements of the computer science course to make room.
Option 4.	Mathematics and Computing departments work together to teach parts of data science in each subject area, and it is assessed as a separate qualification.
Option 5.	Data science is taught in a cross-curricular way to use the 'domains' of other subject areas.

There are many issues around each of these different options. An ideal scenario would be for mathematics and computing teachers to work together so that students take a qualification in data science that comprises both elements (Option 4). However this would require more negotiation between departments than is normally possible in busy, compartmentalized schools. Option 5 would also be appealing as students could look at topics in geography and science for example and apply data science techniques to understand these domains better. This scenario would require significant teacher training. It is clear that there are no obvious answers to how to integrate data science as an assessable subject but it is an issue that will need to be addressed in any consideration of mainstream data science education.

DISCUSSION

Data literacy and data science: breadth and balance in the computing curriculum

In England a new curriculum was introduced in 2014 and new computer science qualifications introduced in 2015 and 2106 for students at age 18 and 16 respectively. We divide Computing into three *strands*: computer science, information technology, and digital literacy, although the GCSE and A-Level only currently focus on computer science. The focus on data in the computing curriculum is largely around data representation (e.g., binary numbers) with limited focus on data collection and analysis. The direction of travel is rather in the wrong direction as there had been more emphasis on data in the previous ICT curriculum.

This is unfortunate as data literacy and data science in the curriculum can address the issue of breadth and balance discussed earlier. Data literacy is important for digital literacy. Working with software to analyze data develops information technology skills. In developing computational models and understanding more complex algorithms, we bring in computer science and thus the topic of data beautifully spans all three strands.

There is a sense that the focus is gradually moving towards the importance of data science and data literacy in the curriculum. Some enthusiasts are beginning to realize how engaging data science can be in school and incorporate it into their lessons or set up teacher training sessions on this topic. This activity is reminiscent of the 'early adopter' behavior often observed around three or four years before a larger-scale change.

Computational thinking: conflicting approaches

The Computing curriculum in England has computational thinking at its core and the focus is on five particular computational thinking concepts, which are algorithms, abstraction, decomposition, generalization, and evaluation (CAS 2015). In contrast, the *nine* aspects of computational thinking proposed by CSTA in USA were outlined in 2011 as follows:

- data collection
- data analysis
- data representation
- problem decomposition
- abstraction
- algorithms
- automation
- parallelization
- simulation

(Barr & Stevenson 2011)

It can be seen that data is a significant part of the understanding of computational thinking described in the CSTA/ISTE guidelines (Barr & Stevenson, 2011) but is missing from that implemented in the curriculum in England. This does not bode well for data literacy in England and is something that may need to be reviewed in future iterations of the curriculum. Simulation is also included which involves developing and understanding models; being able to create data models is a key skill both in computer science and data science.

In addition, Barr & Stevenson (2011) suggest some activities that could be undertaken in the areas of data collection and data analysis, as shown in Table 1 (on the following page).

These example activities are just a tip of the iceberg when it comes to the range of lessons that could be integrated into the curriculum to support data literacy and a focus on data in different subject areas. It indicates that it is not too difficult to incorporate foundational data literacy in mandatory computing (or other subject areas) and that these skills could form a foundation for more technical data science skills in elective computer science or mathematics courses running from 14–16 and 16–18. School structures do not lend themselves well to subjects working together,

but an emphasis on project work across two or more subject areas would make this a viable prospect.

Table 1: Activities suggested for data-related computational thinking skills (adapted from Barr & Stevenson 2011, p.52)

Computational thinking skill	Computer science	Mathematics	Science	Social studies
Data collection	Find a data source for a problem area	Find a data source for a problem area, for example, flipping coins or throwing dice	Collect data from an experiment	Study battle statistics or population data
Data analysis	Write a program to do basic statistical calculations on a set of data	Count occurrences of flips, dice throws and analyzing results	Analyze data from an experiment	Identify trends in data from statistics

CONCLUSION

In this extended abstract, data science in school has been examined from the perspective of a school computing curriculum. This has been illustrated using the example of the curriculum in England where computing exists as a subject from 5–18; this may seem artificial to those countries not in this situation. However, many countries are starting to introduce computing as a mandatory part of the school curriculum (e.g., Japan and Finland). The implementation of the computing curriculum in England has not left much room for teachers to develop lessons around data and this may be an area where improvements could be made in the near future. The recommendation is that ensuring a focus on data literacy in the mandatory years (age 5–14) will leave the door open to include data science as an elective in the later school years.

REFERENCES

Barr, V. & Stephenson, C. (2011). Bringing computational thinking to K–12. *ACM Inroads, 2***,** 48.

Cao, L. (2017). Data science: Challenges and directions. *Communications of the ACM*, *60*(8), 59–68.

Computing At School (CAS) (2015). *Computational thinking: A guide for teachers.* CAS Report. Available at:http://community.computingatschool.org.uk/files/8550/original.pdf

Pybus, J, Coté, M., & Blanke, T. (2015). Hacking the social life of big data. *Big Data & Society*, *2*(2). DOI: 10.1177/2053951715616649

Oceans of Data Institute. (2015). *Building global interest in data literacy: a dialogue*. Waltham, MA: Educational Development Center. Retrieved from http://oceansofdata.org/our-work/building-global-interest-data-literacy-dialogue-workshop-report

The Royal Society. (2017a). *Machine learning: the power and promise of computers that learn by example*. April 2017 DES4702, ISBN: 978-1-78252-259-1. Retrieved from https://royalsociety.org/topics-policy/projects/machine-learning/

The Royal Society. (2017b). *After the reboot: Computing education in UK Schools.* Retrieved from https://royalsociety.org/topics-policy/projects/computing-education/

Wolff, A., Montaner, J.J.C., & Kortuem, G. (2016). Urban Data in the primary classroom: bringing data literacy to the UK curriculum. *The Journal of Community Informatics*, *12*(3).

6 Summarizing perspectives from the discussants

TEACHING DATA SCIENCE AT THE HIGH-SCHOOL AND BEYOND: REFLECTIONS ON GOALS, DILEMMAS, AND COURSE DESIGN

Iddo Gal
Department of Human Services
University of Haifa, Israel
iddo@research.haifa.ac.il

This brief paper reports some reflections regarding the teaching of data science at the high-school level. The paper highlights the heterogeneity of motivations (why) and goals (what) in this field, and points to some dilemmas and tensions that course and curriculum designers should consider.

INTRODUCTION

This paper reports on my reflections regarding the teaching of data science at the high-school level, based on my role as a discussant at the Data Science Symposium held in November 2017 at the University of Paderborn, Germany. The symposium brought together experts and developers from three key communities involved in research and teaching related to data science, i.e., statistics, computer science, and mathematics (and from mathematics education and statistics education), as well as from service providers, users, and funders interested in data science.

As a discussant, I was asked to summarize key impressions from three days of joint work that involved presentations, plenary discussions, and small-group forums. The list below provides some highlights, written in a way that hopefully will provide food for thought for the many stakeholders in this emerging area.

RATIONALE: WHY TEACH/LEARN DATA SCIENCE?

For me, four different motivations or arguments came up at the symposium, for *why* we need to teach data science at the school level, to all students. It is important to note that no speaker mentioned all four arguments listed below, hence listing all of them seems useful:

Data science is an integral part of civic literacy and digital literacy in the 21st century.

Data science fits into and promotes the general goal of schooling, which is to prepare informed citizens. That is, knowledge about data science is part of "what every citizen needs to know or consider." This notion relates to different life roles, i.e., to what citizens need to know as recipients or users of public services, and as users and readers of information in the media; and to what consumers need to know as part of using commercial services (e.g., the need to know about the use of algorithms and big data by service companies and digital platforms), etc.

Data science is needed for workforce preparation in the 21st century.

Data science should be learned because it is part of what the future workforce should know, i.e., its learning is motivated by economic or pragmatic considerations related to "human capital" and to new types of competencies that are required of all workers, managers, and business owners in the age of big data, the internet of things, etc.

Data science is a motivator for learning STEM.

Data science does not "look like" regular mathematics or statistics. Testimonials from programs where data science is taught in high schools suggest that it conveys a sense of innovation to students and relates to experiences and services that young people (i.e., millennials or Generation Z) engage with in various life roles in the digital age. Thus, data science is an attractive topic that can enhance the interest of a wider range of students in learning STEM topics, it can improve attitudes towards science, increase the status of learning STEM topics, etc. Also, data science offers alternative pathways to learn or certify STEM skills for populations that traditionally do not learn advanced mathematical topics, thus it can improve access of minority populations to STEM topics and foster equity in learning STEM, which is seen as a gateway to good or rewarding careers in the information age.

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

Data science enhances the content of what has been traditionally included in STEM.

While the third point, above, related to data science as a motivator, the current argument relates more to the content of learning. The argument here is that understanding data science deepens learning, or expands learning into new subtopics in mathematics, statistics, or computer science and related technology topics that students otherwise may not engage with or that may seem too advanced for many students. Among other things, data science requires acquisition and activation of many big ideas and general concepts related to science, technology and mathematics and statistics (e.g., the use of models and modeling of data, prediction, uncertainty, etc.); it also requires that students develop some understanding of programming, heuristics, ideas related to AI, and the like.

GOALS & OBJECTIVES FOR LEARNING DATA SCIENCE: WHAT TO TEACH?

Beyond the four general arguments for teaching data science, another level of analysis pertains to the specific goals or objectives associated with teaching and learning data science. In other words, the focus now is not on *why* to learn, but *what* needs to be learned. Here, my summary of the various talks and discussions at the symposium, and my own work, points to complementary but separate views on goals and objectives. As before, different speakers each raised only some of the views listed below about aims and objectives, but never all of them under one roof, hence it is valuable to summarize a full list of all views. This list also helps to understand some of pressures, tensions and dilemmas, discussed later on in the closing Discussion section.

Aims related to CONTENT.

Not surprisingly, speakers from the communities most directly involved in developing data science, i.e., statistics, computer science, mathematics, and statistics and mathematics education, as well as from clients of data science (e.g., official statistics, service management, industrial manufacturing), each emphasized a somewhat different set of learning objectives. In other words, each community requires or advocates for more depth or expansion of coverage of *different* subject matter, beyond what is traditionally learned in STEM topics. (Details can be found in the other papers in the conference proceedings). Here are some examples for aims related to new or deeper content that is seen essential for a data science curriculum:

- *Mathematics:* Speakers advocated for deeper or broader learning of the use of models and modeling of data, algorithms and formulae, and other topics.
- *Statistics* (and official statistics): Speakers pushed for learning more about regression, clustering, sampling, uncertainty and probability, administrative data sources, etc.
- *Computer science and data science*: Speakers from these areas emphasized the need to learn more about algorithms, data structures, programming languages, methods for analysis of big data, data mining, machine learning, predictive analytics, and related topics.
- *Service science, management*: Speakers representing various users of data science emphasized more content related to client needs, such as concepts and knowledge related to increase in value or efficiency of processes, or improved outcomes in terms of, e.g., profits, marketing, industrial production, citizen engagement (e.g., in smart cities), service quality, customer satisfaction, etc.

Aims related to LEARNING PROCESS.

While less pronounced, some speakers focused on the potential of learning data science to promote learning goals that are sometimes difficult to achieve in regular classes in science, technology, and computer science, and mathematics and statistics, especially as they are taught or learned at the *high-school* level. Key areas are (but not limited to):

- Data science enables students to experience *inquiry-based, problem-based learning.*
- Data science enables development of *teamwork* and *collaboration* skills (in part via technology).

- Data science enables students to work on real-world data, and real problems representing needs of client organizations or communities, hence it is a vehicle for *linking students to contexts and broader social and economic problems and processes*.

The above list implies that the methods and organization of instruction in data science should allocate time for many new knowledge and skill areas that are dictated by learning processes as much as by desired content. Time and energy should be devoted to planning team-based work and communication processes, bringing in context experts (e.g., from client areas), learning to deal with messy real data and not with cleaned or fabricated data, and so forth.

Aims related to adding value for TEACHERS and STUDENTS.

While the ideas regarding content and process outlined above appear most prominent, some participants mentioned fuzzy goals related to the value added by learning data science. In this view, data science does not fit the traditional boundaries between mathematics, statistics, and computer science, but rather extends beyond them, yet also requires integration between topics or disciplines that from the point of view of many students are separate from each other. Hence, the umbrella of data science creates a new learning context, beyond traditional STEM labels, where students can acquire new STEM-related knowledge, regardless of whether they are "advanced" or "average" or "weak." In this view, it does not really matter what new content teachers or students learn in data science, but just the fact that they will learn *some* new material that otherwise they would not have acquired or be interested in acquiring, thus promoting general aspirations to improve STEM education, at any level of instruction.

DISCUSSION: PRESSURES, DILEMMAS, AND FUTURE CHALLENGES.

Taking a broad look at the arguments and views above, they may seem to complement each other, but each focuses on a different issue. The multiplicity of views about *why* and *what* to teach regarding data science overall creates pressures and dilemmas that did not seem obvious to many participants in the symposium.

The four general arguments for *why* teach data science form two clusters: The first two pertain to external considerations, that is, how knowledge of data science is related to demands or changes in the world in which we live. The last two are more internal to school (or university) systems and relate to considerations or factors associated with the motivation for learning STEM topics and who studies what within STEM.

It is important to reiterate that the arguments listed above reflect different ideas raised by speakers and participants at the conference (see other papers in these proceedings). Yet, each speaker mentioned only some of the arguments, and all of them were not raised under one roof, hence there is merit in looking at them as a coherent framework for motivating and promoting the inclusion of data science in school curricula, or creating new elective courses in data science that will be offered alongside regular topics.

At the same time, the arguments and views summarized above in sections on learning process and added value for teachers and students certainly push and pull in different directions. If we think of a basic one-semester course in data science (e.g., 2 weekly hours for 13–14 weeks), it does not seem possible to cover all the new ideas and subtopics listed by each one of the disciplines as outlined above under "Aims related to CONTENT." The starting point for curriculum design will vary greatly depending on assumptions we can make about the students' level of knowledge and confidence in mathematics, statistics, and computer science. Further, different learning trajectories can be envisioned for students, depending on whether they start at a "low," "average," or "advanced" levels in these or related areas.

Likewise, teachers' *content* knowledge and *pedagogical* content knowledge will vary greatly in each of the areas related to data science. From this perspective, the design of a data science curriculum for the high-school context seems a more challenging endeavor compared to putting together a full degree program at the university level, since multi-faceted content knowledge in multiple disciplines is needed of a single teacher. We can expect many "sparse expertise" issues among teachers that will require much investment in professional development and support materials.

Overall, the range and number of motivations, aims and objectives, and possible working assumptions about students and teachers that were sketched above will require careful decisions and some tradeoffs from course designers and curriculum planners.

DATA SCIENCE AS A SCHOOL SUBJECT IN SECONDARY EDUCATION FROM THE PERSPECTIVE OF COMPUTER SCIENCE EDUCATION

Johannes Magenheim
Institute of Informatics
University of Paderborn
jsm@uni-paderborn.de

This extended abstract summarizes the results of the symposium's discussions on Data Science (DS) as a school subject in higher secondary schools from the perspective of Computer Science Education (CSE). The emphasis of the presentation is: The rationale (why should students learn about DS?), the aims and objectives (which objectives should students learn, which competencies should students achieve?), the content (What are most relevant learnings?) and finally, implementation strategies of this subject area in a school curriculum and existing constraints.

DATA SCIENCE IN THE CONTEXT OF SOCIO-TECHNOLOGICAL DEVELOPMENTS

An essential characteristic of the application of information and communication technologies (ICT) in different areas of society is the mass-generation, automated processing, collection, storage, and distribution of data. This data is needed, e.g., for the control of processes, but also can be generated as results and side-products of technical processes, human-computer interactions, or derived from social interactions of humans. Fostering an understanding of how these large data volumes are handled in the information society and how they are processed by networked IT systems is an important objective of informatics education or computer science education (CSE). The creation of a critical awareness towards accessibility and protection of this data is regarded as an important contribution of computer science education to general education.

The conscious and reflective handling of data—their own data as well as external data—requires that students acquire basic knowledge in the areas of security and privacy, and are aware of the opportunities and risks involved in the automated processing of big data. In addition to the basic understanding of technical concepts for processing big data, this also includes a certain insight into common application scenarios of data processing and the social consequences that may result from them, such as, profiling, data retention, weblining, social exclusion, and stigmatization. In this respect, the problem of formal and technical processing and the legal and social handling of big data is, on the one hand, a central object of Data Science (DS) and at the same time an important area in the intersection between DS and Computer Science (CS).

IT systems that process large amounts of data, such as search engines or social software, would not be technically feasible without the underlying efficient algorithms. In this respect, algorithms have always played an important role in CS. To understand the processing of large amounts of data, it is not enough to understand common algorithms and data structures. It is also, for example, about the handling of uncertain and incomplete information. This is where methods from DS come into play, such as combinatorics, probability calculus, and stochastics, as well as descriptive and inferential statistics.

Dealing with big data requires processing concepts that exceed CS's core methodological areas and lie within the overlapping area of both scientific disciplines. For this reason, precisely these subject areas should also be the subject of a corresponding learning area at school, which can contain not only computer science and mathematical references but also socio-scientific economic and ecological subject areas.

OBJECTIVES AND COMPETENCIES ACQUISITION

What contribution can CSE make to a DS learning area at secondary schools? Is it possible that the objectives of CSE can be fully aligned with the demands on the educational system, resulting from the social developments described above? An analysis of existing CSE curricula shows that although CSE can make an essential contribution to Data Science Education (DSE), full integration of DSE in CSE is hardly possible. There are objectives and areas of demand in DSE that

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

can be easily reconciled with CSE-objectives and the expected students' acquisition of CSE-competencies. On the other hand, there are also objectives that go beyond the competence-areas that CSE has been striving for so far and, moreover, primarily include mathematical areas of competence. Besides, CSE includes essential objectives that are not at all covered by DSE. These interrelationships between CSE and DSE will be explained in more detail in the following.

Almost all CSE curricula for Upper and Lower Secondary Education in the German States are based on the Educational Standards for Computer Science Education (GI 2008, 2016). These recommendations for CSE standards are based on the NCTM (2000) competence structure model, which is also the basis for the educational standards of mathematics in Germany. In the NCTM and GI recommendations, competencies in the sense of Weinert (2001) are regarded as observable and executable action plans (processes) which are contextualized in the respective content area. Therefore, the competencies students should achieve can be described as the result of a combination of content and process aspects. Both components are not distinctly separated from each other.

The GI educational standards describe five process-related and five content-related competence components:

Process-related Competence Components	Content-related Competence Components
P1: Modeling and Implementing	C1: Information and Data
P2: Evaluating and Reasoning	C2: Algorithms
P3: Structuring and Cross-Linking	C3: Language and Automatons
P4: Communicating and Co-Operating	C4: Informatics Systems
P5: Represent (Visualize) and Interpret	C5: Informatics, Humans and Society

Recent didactic research argues for the introduction of an additional process area P0 'Exploring and Interacting' (Bergner et al. 2017) and provides a didactically substantiated foundation for this process area (Schulte et al. 2017).

This scheme allows deriving a whole series of competence expectations, which contain a reliable reference to important DSE objectives and contents. Here are some examples, defined on a general level:

- (P1/C1) - Ability to model and clean large datasets to clarify specific questions;
- (P1/C1/C4) - Ability to analyze data using computer systems to explain queries or extract information contained in the data;
- (P2/C2/C2/C4) - Ability to appropriately interpret the results of 'data mining' with regard to a given hypothesis
- (P3/C5) - Ability to recognize the societal implications of networked data storage in informatics systems;
- (P4/C4) - Ability to successfully implement large projects using data mining methods by means of suitable IT-systems in the context of cooperative work processes;
- (P5/C4/C5) - Communicate and interpret the results of data mining projects by means of appropriate visualizations of the results with reference to their social context;
- (P0/C2/C4) - Ability to explore the functionality of data mining software tools in terms of their appropriate application and the understanding of the algorithms and calculation methods they use.

To enable students to acquire competencies during CS-lessons, in a sense described above, we have to address appropriate context-specific questions on how to deal with big data in a methodically appropriate manner (see below).

In national and international CSE curricula, further concretizations have been carried out to implement competence-areas concerning the processing of big data. For example, the German Informatics Society(GI) issued computer science education standards for higher secondary education in Germany on a higher requirement level (GI 2016). According to these recommendations, the students should be able to…

- "use, model, and implement operations on complex data structures"
- "develop a database for a real-world problem that involves complex relationships"
- "analyze communication and data storage in networked systems and assess them also from the view of data protection and data security" (GI 2016, p. 10; author's translation).

Here, the processing of more massive amounts of data is mainly discussed in relation to structured data within the framework and methods of databases, but also unstructured data sets in distributed systems can be topics of CS-lessons.

More recent international CSE curricula propose much more explicitly to deal with big data by applying specific analytical methods within computer science education:

- "Use data analysis tools and techniques to identify patterns in data representing complex systems. For example, identify trends in a dataset representing social media interactions, movie reviews, or shopping patterns"
- "Select data collection tools and techniques to generate data sets that support a claim or
- communicate information."
- "Evaluate the ability of models and simulations to test and support the refinement of hypotheses" (CSTA 2017, p17)

Students shall acquire these skills according to the CSTA K–12 recommendations within the data-related topic-areas 'Collection, Visualization & Transformation,' 'Hypotheses,' and 'Inference & Models and Algorithms.'

In a conceptual description of a modern CSE curriculum in the UK, the Royal Society also emphasizes the significance of big data: "The opportunities provided by new computing curricula coupled with advances in technologies and analytical tools with which to mine big datasets, and the increasingly interdisciplinary nature of educational research, offer enormous scope for advancing computing teaching and learning" (Royal Society 2017, p. 96).

The competency-recommendations of newer CSE curricula, concerning the areas of information and data, are explicitly related to operations with large, partially unstructured data. This reference reveals that it is necessary for the successful implementation of these competence requirements, that students acquire required mathematical skills in the fields of inferential statistics, combinatorics, probability calculus, and statistics. These necessary mathematical skills go beyond the algorithms, customarily dealt with in CSE, and are not part of a traditional CSE curriculum. In this respect, these fields of competence and topics transcend those addressed by classical CSE curricula.

On the other hand, CSE curricula comprise essential areas of competence that are not in the focus of DSE, for example, exploring and understanding informatics systems concerning their externally visible function and their internal structure. The results of that kind of exploration will enable students to enhance their informatics systems and to design and to develop new ones. In doing so, the respective socio-technical application context of the informatics systems and the capabilities and the potential interests of the users are to be taken into account. These kinds of questions and fields of competence are the focus of various CSE approaches that regard informatics systems as socio-technical systems. According to this view, informatics systems are characterized by the duality of internal structure and externally visible functions. Students may learn about this duality by acting and learning along with a didactically interwoven learning path, considered as an action cycle of system exploration and system design. Questions about the systems' usability and the GUI-construction are also subject to exploration and construction processes (Schulte et al. 2017). It is then not about the use of software as an analytical tool for the needs of data analysis but rather about designing and creating an informatics system and in this way understanding how such tools internally work. Learning about tools and not learning with tools is a focus of CSE learning processes. This difference also demonstrates an important task of CSE: Illuminate the black box 'informatics system' a little bit. All in all, these areas of competence show common features as well as apparent differences between CSE and DSE requirements for a school curriculum.

In the field of non-cognitive competencies, which according to Weinert includes a volitional and a motivational competence-dimension, CSE and DSE reveal a high degree of alignment with school curricula. Especially during interdisciplinary computer science projects about essential aspects of the students' real lives, when collecting, analyzing, evaluating, and presenting data, the students' ability to work communicatively and cooperatively in a team can be fostered. If the topics of CSE-lessons address real-world problems with vital relevance to the students' real life, there are also good possibilities to promote students' motivation and self-efficacy with regard to operating and understanding informatics systems as well as data processing and information retrieval.

The KMK's strategy paper on 'Education in the digital world' (KMK 2016, p. 15) also addresses the promotion of non-cognitive competences and the use of IT systems as analytical tools. This section describes areas of competence which have a high affinity for handling big data and which can be very well implemented in CSE curricula:

- Searching, processing, and storing (search and filtering / assessing and validating / saving and retrieval);
- Communicating and cooperating (interacting / sharing);
- Production and presentation (developing and producing / further processing and integrating / observing legal regulations);
- Protecting and acting safely (acting securely in digital environments / protecting personal data and obeying privacy issues);
- Problem solving and action (solving technical problems / using tools appropriately to the tasks / identifying deficits and looking for solutions / using digital tools and media for learning, working and problem solving / identifying and formulating algorithms);
- Analyzing and reflecting (analysis and evaluation of media / understand and reflect on media in the digital world (analysis and evaluation of the benefits and risks of business activities and services on the internet; recognition, analysis and reflection of the potential of digitalization in terms of social integration and participation; [ta]).

CONTENT, METHODS, AND TOOLS

According to the remarks above, the topic of big data can help motivate learning processes for students during CSE lessons. The use of real world data can ensure an affective relation of the students to this topic and can stimulate their motivation. Therefore, the use of publicly accessible data sources such as ProCivicStat in the EU (ProCivicStat 2018) or, for German-specific data, the 'Statistisches Bundesamt' (Statistisches Bundesamt 2018).

The concept of participatory sensing offers a further possibility to collect and evaluate data in a way that motivates learners. By using mobile devices and their embedded sensors, data can be automatically recorded by many people on time- and site-specific basis and simultaneously forwarded to a central server.

The data are then available on the server and can be used for assessment regarding different 'research' questions. Since the data is generated by many people involved, this type of data collection is also called *crowdsourcing*. Data from the standard sensors of different mobile devices can be automatically collected (e.g., GPS data, altimeters, medical data of the users...). Besides, data from users of mobile devices can also be deliberately generated and transmitted to the server (photos, videos, sound recordings, etc.).

In this way, depending on the research question, it is possible to record environmental values automatically, or participants of a project can additionally submit data to the server (e.g., weather data, data on water quality, traffic data, etc.). This form of data collection offers a large number of exciting possibilities for the evaluation of various course topics and interdisciplinary teaching projects of various sizes in which learning groups from different schools can be involved. It is even possible to organize international projects in this way.

From an informatics perspective, which transcends and enhances the DSE view on a topic, the mobile devices with their sensors and the applied client-server concept that is used for data transmission, can be regarded as socio-technical informatics systems. These informatics systems can be analyzed and evaluated during CSE lessons about the systems' technical and social aspects

and, if necessary, the students can modify and re-design the informatics systems for specific needs of their intended evaluation.

Depending on the problem to be examined, appropriate data sets must be selected here. In contrast to the databases traditionally treated in computer science classes, which are usually available as tables and relations, these data are often semi- or un-structured, sometimes from different sources, and have to be adjusted according to the given assignment. Methods of 'cleaning data,' 'wrangling data,' or 'munging data' have to be applied as a first step of data analysis. These procedures of data transformation from raw data into an appropriate form that is more suitable for the analysis of the given assignment, can be regarded as a data modeling process, which is quite common in computer science classes even for the modeling of informatics systems.

Afterwards, these data can be further processed with methods of combinatorics, probability calculation, and inferential statistics. Suitable software tools for these calculation methods are available, such as

- CODAP (https://codap.concord.org),
- Fathom (https://fathom.concord.org),
- TinkerPlots (https://www.tinkerplots.com) or
- inZight (https://www.stat.auckland.ac.nz/~wild/iNZight/index.php).

The adequate spatial representation of the results, concerning the question raised at the beginning of the data analysis, can also be regarded as a contribution to CSE lessons because the results of informatics system development and the systems' performance should also be presented in a manner comprehensible to the user.

As an alternative to the use of analytic tools, students can develop their scripts for the necessary calculations, which can then be processed by generic tools such as R or Python. In the latter case, the calculation procedures and the functioning of the applied informatics system remain transparent. However, this learning strategy requires basic knowledge of both stochastics and programming on the part of the students. But the necessary knowledge can be acquired integratively by the students in an appropriately designed CSE course. On the other hand, it seems relatively difficult to explore and enhance an application tool for data analysis, as mentioned above, with methods of deconstruction and re-engineering. Also, the construction of that kind of DSE-software-tools seems to be out of scope for most of the students in CSE-classes. Thus, it is supposed that the content areas of system design and system development will not be in the focus of an integrated CSE/DSE didactical approach.

Nevertheless, programming can also take into account other specialist IT concepts, such as the programming of neural networks. Especially in the context of machine learning and deep learning, the modeling and programming of neural networks open up another interesting subject area for computer science teaching. "Deep learning is a specific subfield of machine learning: a new take on learning representations from data that puts an emphasis on learning successive layers of increasingly meaningful representations." (Chollet 2018, p. 8).

On the one hand, you could use a framework or shell of a neural network to brief the neural network with existing data in a learning mode. The available data could then be used to identify patterns or forecast future developments. On the other hand, a corresponding neural network with different layers could also be successively modeled and implemented with Python. In this case, it could also be used for prediction purposes or pattern recognition, possibly even from incomplete information. The concepts of the mere usage of a tool for data analysis and the design of such a tool with a subsequent application represent two different approaches to DS within CSE.

However, an integrative didactical concept, which concentrates on the programming of scripts for data evaluation, seems to offer an acceptable way to teach the content and methods of CS and DS appropriately in a CS course. Such an approach can enable students to overcome the boundaries from pure math and statistics to modeling, system design, and 'computational thinking.'

Examples of such course concepts are listed below:

- The evaluation of microblogs (tweets) regarding specific topics (hashtags) and the time course of their occurrence. For this purpose, interfaces can be used to access the raw data

(e.g., https://developer.twitter.com/en/docs) or apply already existing analysis tools (e.g., https://www.talkwalker.com/de/blog/7-kostenlose-twitter-analyse-tools-empfohlen-von-experten).

- Content- and time-related analysis of communication (via different media) in accessible parts of social networks (AAN: Artifact Actor Network Analysis; e.g., Riss et. al. 2011);
- Analysis of persons (groups) and documents with regard to social and content-related tagging;
- Forecast of sports results, development of stocks based on past performance;
- Conducting election forecasts based on previous data and surveys conducted by the students
- (e.g. http://www.bpb.de/lernen/grafstat/grafstat-bundestagswahl-2013/144674/bundestagswahl-wahlanalyse-und-wahlprognose);
- Evaluation of data collected by students with mobile devices concerning specific questions (e.g., scientific field excursions such as in Schaal et al., 2011 and Holtdorf 2014);
- Data analysis in the context of learning analytics and usability (e.g., assessment of eye tracking data);
- Evaluation of GPS movement data of persons, deliberately collected with mobile devices (e. g. smartwatch, bicycle computer, pedometer...) with the associated medical and physiological data;
- Evaluation of traffic flow data (e.g., applying pattern recognition from webcams);
- Analysis of text documents about semantic similarity (or co-authoring, plagiarism);
- Pattern recognition within pictures and videos (e.g. images of drones, NAO-face recognition: see Schäfer & Schlee (2018) etc.
- Analysis of weather data (https://openweathermap.org);
- International projects concerning astronomical and physical satellite data (e.g., My-NASADATA: https://mynasadata.larc.nasa.gov/students/).

IMPLEMENTATION STRATEGIES AND CONSTRAINTS

The question arises how DSE can be integrated into existing school curricula. Because of a large number of subjects and learning areas and full timetables, the introduction of an additional subject at German secondary schools appears to be less promising. Thus, despite years of sustained efforts, it has not yet been possible to establish the discipline of computer science education as an integral part of the school curriculum at all levels. Only in some federal states in Germany, are binding regulations for mandatory CSE regarding certain types of schools (usually ‘Gymnasium’) and grades. Initiating another initiative for the implementation of a DS- subject would be very difficult.

The introduction of DS as an optional interdisciplinary learning area is also problematic, which is illustrated by a look at the recent history of introducing learning areas into German school curricula. In particular, the introduction of basic informatics education and integrative media literacy provide dissuasive examples. They were planned as interdisciplinary curricular learning areas that, depending on the didactic concept of the respective federal state, were integrated into the traditional school subjects. In practice, however, the specific curricular implementation of the concepts failed due to inappropriately qualified teachers and their lack of willingness and ability to teach a particular part of their subject lessons in an integrative way, by considering ICT-issues. The lack of availability of teaching materials also proved to be a significant obstacle in the introductory phase of the integrative learning areas.

On the other hand, the previous remarks have shown that there are many links between DSE objectives and content and those of CSE. To teach these DS topics in a motivating and understandable way for the students, requires some mathematical knowledge. It is therefore advisable to integrate DS topics into the curricula of both subjects and to align the sequence of methods and issues. Besides, it is possible to acquire the necessary mathematical knowledge in an integrated manner, within the scope of DS topics in computer science lessons. The alignment of the curricula between mathematics and computer science should also start at the beginning secondary level, where essential basics of stochastics could be taught in math. On the other hand, in informatics, the

scope of the topic of databases could be reduced and partly shifted to compulsory elective courses in grades 9 and 10. The logic required for understanding database operations and the relational algebra can also be taught in the first year of secondary school. It is also possible, for example, to make iterative predictions for future developments from existing data using spreadsheet-based algorithms and other numerical calculations. This could also be a good place to introduce inferential statistical methods.

In this way, we take a first step towards the description of trajectories that involve a gradual increase in the aspiration level and the complexity of the DS topics to be covered in computer science lessons. In the field of mathematics, this step could involve a transition from descriptive to analytical methods that enable students to describe samples and predict future developments. In computer science, this phase can include the transition from structured data sets (databases) to extensive unstructured data collections and corresponding automated analysis procedures.

In the CSE lessons of the higher secondary level, DS topics and methods can be arranged as described in the previous section. Besides, there is the possibility to conduct CS teaching projects that contain interdisciplinary references. Classroom projects in the subject of computer science also offer students the opportunity to acquire non-cognitive skills in addition to subject-related knowledge, as is also necessary for collaborative work with big data. Such teaching projects can also be organized for a whole semester, e.g., as a project course, as intended in the CSE curricula of some Federal states in Germany.

An essential element for the successful implementation of DS in school curricula is well-founded teacher training, independent of the chosen implementation strategy. From today's point of view, it seems most useful to focus first on mathematics teachers and computer science teachers, and then gradually expand the range of training to interested teachers of other subjects. Furthermore, an accompanying formative evaluation, as an essential control element for the experiences gained in practice, is another critical building block of a successful implementation strategy.

REFERENCES

Chollet, F. (2018). *Deep learning with Python*. Shelter Island (NY): Manning Publications.

CSTA/ACM (Ed.) (2017). *K–12 computer science standards*, Revised 2017. http://www.csteachers.org/page/standards

Bergner, N., Köster H., Magenheim, J., Müller, K., Romeike, R., Schulte, C., & Schroeder, U. (2017). Zieldimensionen informatischer Bildung im Elementar- und Primarbereich. In Stiftung Haus der kleinen Forscher (Hrsg.), *Frühe informatische Bildung – Ziele und Gelingensbedingungen für den Elementar- und Primarbereich*. Berlin 2017

GI (Gesellschaft für Informatik) (Hrsg.) (2008). *Grundsätze und Standards für die Informatik in der Schule: Bildungsstandards Informatik für die Sekundarstufe I, 2008*. http://www.informatikstandards.de

GI (Gesellschaft für Informatik) (Hrsg.) (2016). *Bildungsstandards Informatik für die Sekundarstufe II, Empfehlungen der Gesellschaft für Informatik e.V. erarbeitet vom Arbeitskreis »Bildungsstandards SII«* http://www.informatikstandards.de/docs/Bildungsstandards_SII.pdf

Holtdorf, J. (2014). *Konzept und Implementierung eines Smartphone basierten Systems zur Unterstützung bei Feldversuchen im Schulunterricht*, Masterthesis, CSE Research Group, University of Paderborn.

Kultusministerkonferenz KMK (2016) (Ed.). *Strategie der Kultusministerkonferenz „Bildung in der digitalen Welt."* http://bit.ly/2vCRrUP

Schäfer, D., & Schlee, H. (2018) *Bildverarbeitung; NAO Gesichtserkennung*. http://bit.ly/2vBNRKS (PowerPoint presentation on facial recognition.)

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*.

ProCivicStat. (2018). Promoting civic engagement via explorations of evidence http://community.dur.ac.uk/procivic.stat/

Riss, U., Magenheim, J., Reinhardt, W., Nelkner, T., & Hinkelmann, K. (2011). Added value of sociofact analysis for business agility. In *AAI Spring Symposium Series 2011*. Retrieved from http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2444

Schaal, S., Matt, M., Bullinger, M., Fauth, S., & Dingler, C. (2011). *Forschend-entdeckendes Lernen im naturwissenschaftlichen Unterricht mit mobilen Technologien*; http://bit.ly/2vxhHjm

Schulte, C., Magenheim, J., Müller, K., & Budde, L. (2017). The design and exploration cycle as research and development framework in computing education. In *Global Engineering Education Conference (EDUCON), 2017 IEEE* (pp. 867–876). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/7942950/

Statistisches Bundesamt (2018). https://www.destatis.de/DE/Startseite.html

The Royal Society (Ed.) (2017), *After the reboot: Computing education in UK schools*, Issued November 2017. Retrieved from http://royalsociety.org/computing-education

Weinert, F. E. (2001), *Leistungsmessungen in Schulen*. Weinheim: Belt

TOOLS, BEST PRACTICES, AND RESEARCH-BASED REMINDERS

Andee Rubin and Tim Erickson
TERC, Cambridge, MA, USA; Epistemological Engineering, Oakland, CA, USA
andee_rubin@terc.edu; eepsmedia@gmail.com

This extended abstract presents a broad overview of the symposium's conversation on two topics in data science education: "Tools" and "Best Practices." In addition, we will reflect on what research might have to say about data science education in general, and on those topics in particular.

TOOLS

A number of presentations and discussions directly addressed or evoked the question of what tools to use in data science education. A tension emerged between what we might call "professional" tools (such as R or Python, often scaffolded with RStudio or Jupyter Notebook) and "learning" tools (such as inZight, TinkerPlots, Fathom, or CODAP).

Proponents of using professional tools point to a number of advantages:

- Professional tools often support the convenient attachment of metadata.
- A notebook documents your thinking explicitly and sequentially.
- It's easy to apply a solution to a new data set.
- Professional tools work well with truly *large* large data sets.
- Professional tools are used in the workplace, so using them in education prepares students for jobs.

Proponents of learning tools note that:

- Learning tools are accessible; it's easier to get started.
- The user experience is more "tactile" and dynamic.
- An analysis is embedded in a visible configuration rather than expressed abstractly in code.
- They can expose and explain parts of a complex process.
- They are explicitly designed to support learning underlying concepts.

The difference is also sometimes characterized by *coding*. Using R or Python, at some level, involves writing and executing computer code. This is a disadvantage in that code can be abstract and daunting; but it's an advantage in that code elegantly and reliably expresses the steps of an analysis and can be re-executed with different data.

Whatever tools you decide to use, the use of any tool presents challenges for an instructor. When and how should students learn to use it? How much overhead does learning a tool take up? Do students focus so much on the tool that they lose sight of the data science? Or is using a tool actually part of what it means to do data science?

Finally, what access do students have to tools? When we ask this question, we might be thinking simply of accessibility to hardware and software; but we should also consider that students come with a variety of preparatory experiences. Which students are ready to use which tools? We must be careful that, in bringing data science to a broader population of learners, that we do not create equity problems.

One strategy for choosing tools might be a blend of these tools: begin with a learning tool—with its friendlier beginning—then transition to a professional tool when (and if) the learning tool ceases to be effective, given the contexts and problems we are presenting.

An entirely different perspective avoids this learning/professional split. Instead, it asks, when should we choose different tools for different tasks, and when should we choose a tool as a world to live in? The former seems best on the surface: choose tools based on what you're trying to accomplish; but the latter has the advantage of letting us develop a deep understanding of an environment.

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

BEST PRACTICES

The symposium—in presentations and in subsequent discussions—addressed issues of best practices. What pedagogical practices should we use in data science education? What topics should we pursue? What data should we include in tasks we design?

Overall Perspective

As a group, we were quite progressive in our approach. Four overarching principles seemed to shine through:

- The work should be hands-on. Students should actually get to work with the data, and *do* analysis rather than simply hear or read about it.
- Any curriculum and presentation must attend to students' initial understanding. We must be alert to the knowledge and skills students need—but might not have—about data, computational thinking, statistics, and the underlying context of an activity.
- Our approach should illuminate the principles underlying an analysis rather than simply explaining how to do it.
- We should put power into the students' hands. Although it is not always possible or practical, students should be able to choose what problems to investigate, what techniques to use, what data to study, and possibly even what actions to take based on the results of their analysis.

Topics for Analysis

What data will students study? What contexts and issues? In the symposium we heard about and discussed a wide variety of topics. We invite the reader to learn about all of them in the abstracts from the individual presentations. It's worth noting a few overarching ideas here.

First, social issues appeared repeatedly. These included topics in which the student may have a personal stake, for example, the issue of privacy in data. Then we have broader social topics where students bring outside knowledge to the table—data about people and what they do, sometimes as individuals but often in the aggregate. In any social issue, note that for some "hot-button" issues such as race, instructors need good listening and facilitation skills.

Second, some of us, in our teaching of statistics or data analysis, have focused on procedures we might classify as *estimation* or *regression*. In the symposium, a number of presentations and discussions highlighted *classification* situations as characteristic of data science. It seems that some classification tasks and techniques might be at once interesting, accessible (easy to understand what is being asked), and challenging for students. One can imagine combining this with social data, for example, by developing a scheme for deciding which developing countries most urgently need aid.

Finally, a statistics issue: we now have so much data that nearly every difference is significant. Does that mean that inference is dead? No, but we have to rethink the role of inferential statistics in this new world of data science. A good first step will be to notice and promote the role of *informal inference* in formulating data science results. This has the happy consequence of making data science more accessible to students with less experience in formal statistics. As we move forward, we might also want to reconsider and re-prioritize statistics topics. Bayesian thinking, for example, might be more important than it has been in the past.

Working in Teams

A number of presentations described data science *teams* in the workplace, including some well-developed schemes for organizing such group work over time. We must wonder, therefore, how important it is for students to work in teams in a school or university class. There is a lot to be said for the idea. An authentic data science task is often too big for one person. It could be interesting and useful for students to learn a practical schema for a data science team working on a project. And working in teams could be attractive to some students who are otherwise put off by data and technology.

Teamwork in school, of course, has its perils as well. Instructors may need to educate themselves in the great body of knowledge about how to handle groups, ensure fair participation, do authentic assessment, and so forth.

Data Handling in the Curriculum

Traditionally, teachers of data analysis give students clean, pre-digested data sets; the rationale has been that we don't want to spend time on setting up the data—we're focusing on some analytical topic. Yet most definitions of data science stress that handling data—independent of its context or the details of its analysis—is an important part of the work. Perhaps we should finally include various forms of data handling as topics in the curriculum. Here are five overlapping ideas about what we might include:

- Data retrieval. What is it like to go out on the Internet and download some public data? The more extensive the data repository, often the more confusing the interface.
- Data cleaning. When you get the data, it often has properties (for example, using 999 for missing data, or punctuation that confuses the software) that make it unsuitable for further analysis—and you have to fix it. Practitioners often claim that this is a large percentage of their work.
- Data wrangling. Even if the data are clean, they often need to be altered, connected, or related in various ways. Substituting text values for numerical codes, or otherwise connecting multiple data sets are examples of this.
- Data munging. Fully connected and clean data still sometimes need additional work to be fully useful. One example is, suppose you have timecodes for events, and you want to know how many of them happened on which day of the week? You need some understanding of your system's *Date* data type.
- Data formats for archival. If we ever want to use our data again, how should we store it so we can use it easily? How can we store it so it's most accessible to others? How can we include meta-data in the most accessible and convenient way? Note that we have come full circle from retrieval, at the head of this list.

Data Sources

Finally, what data shall we use? Government and industry are creating increasingly large, and often increasingly accessible data sets. There are emerging repositories, but there is no particular agreement on formats at this time.

For our specific community—and we hope for more practitioners as it becomes better known—we have *ProCivicStat*, which explicitly embraces many of the values we espouse here.

Whatever source we find, we want to figure out the best ways to let students choose what they will work on, whether it is data they seek and retrieve themselves or a data set from a selection we give them. If we do that, though, we have to bear two problems in mind: the difficulty of handling any data (retrieval, munging, cleaning, etc.) and the problem of applicability, that is, whether the data truly lets the student experience the data-analysis challenge we're trying to present in our assignments.

RESEARCH-BASED REMINDERS

While data science education has taken on new urgency with the rising importance of data science in the workplace and public forums, we should be sure to remember the results of multiple decades of research into statistics education. Several insights from this body of literature follow:

Proportional reasoning

While proportional reasoning isn't often considered part of the domain of data science, research with high school students using data to create infographics (see Rubin 2018) has highlighted the difficulty some of them have with part/whole relationships. Figuring out what constitutes the "whole" and what "parts" comprise such a whole is not always clear to novices, especially when they are representing data already expressed as percentages, rather than deriving

the percentages from case data. Misleading statements and poor visual representations can result from these kinds of misunderstandings.

Aggregate thinking

Being able to reason about data in the aggregate, taking into account distributional attributes such as shape, center, and variability, can be difficult for novice data analysts. In their initial encounters with data, students tend to focus on individual points or on modal values; it takes experience and practice to know how to balance this legitimate concern for individuals with a broader view of the data.

Uncertainty and probability

Much research has focused on students' (and non-students') struggles with probability. It is possible, however, that with the increasing use of large data sets that are not samples, that issues of sampling variability will fade into the background. But there is still uncertainty in conclusions drawn from data; how do our approaches to teaching about probability need to be modified to take account of this shift?

Geo-spatial data

While statistics education research has put considerable effort into studying students' thinking about data, the emphasis has been on data that is visualized on X-Y plots. Data scientists, however, often use geo-spatial data, visualized on a map. Statistics education researchers have in general paid less attention to the perceptual and cognitive issues that come up in the analysis of spatial data, so we know less about the potential difficulties students may face. Does the notion of "aggregate thinking" apply in the same way to spatial data or does it manifest somewhat differently?

DATA SCIENCE IN SCHOOLS FROM THE PERSPECTIVE OF CONTEXTUAL INFORMATICS

Harald Selke
Heinz Nixdorf Institut
University of Paderborn
hase@uni-paderborn.de

This contribution tries to draw some conclusions from the presentations given and discussions held at the symposium on data science education at school level from a social and cultural perspective, highlighting questions of what the key concepts to be taught are and what relevance data science might have for the world outside school and beyond the individual scientific disciplines involved.

INTRODUCTION

In this short contribution I am trying to summarize my thoughts on the topics addressed in the symposium *Perspectives for data science education at school level* held in late 2017 in Paderborn, Germany. I was invited there not as an expert in data science (which I am not) but rather to discuss the presentations given and the discussions on that symposium from a social and cultural perspective.

Being a computer scientist myself with a background in mathematics, my research interests are in the field that we call contextual informatics, where questions are addressed on the non-mathematical theories of computer science such as the relation of digital technologies to the human mind, the influences of society on computer science, and the effects of computer science on society and the individual. Thus, I was asked to focus on three key aspects:

- The rationale—why should data science be taught in school?
- The aims and objectives—which goals are to be achieved?
- The content—what is it that students in school should be learning about data science?

WHY SHOULD DATA SCIENCE BE TAUGHT IN SCHOOL?

The first question for me to be answered was: What exactly is data science? It obviously draws from methods of statistics as well as from certain fields in computer science, but from my point of view, data science should be more just the application of methods from other fields in order to be considered for school teaching. My first shot at trying to identify the core of data science from what I have learnt at this symposium would be the following attempt of a definition: *In data science, large amounts of data are collected either for different purposes or without a specific purpose. These data are then later scrutinized and in many cases associated with other data by automated or interactive processes.*

This definition attempt, on the one hand, points at one key difference from statistics, where usually the focus is on comparatively small amounts of data that have been collected for a specific purpose to help answer questions that have been stated before the collection of the data. Also, the sources and characteristics of the data are usually defined in advance, and the methods for analyzing the data have been specified.

On the other hand, an important development in computing applications is also visible here: Formerly unrelated data that were residing in different repositories are now being integrated into even larger repositories—which requires new methods of analyzing, visualizing, and using those data. Thus, there is also some potential here for new aspects of computer science that might need to be taught in school.

Tobias Matzner (2018) in his talk mentioned five reasons why data science competencies might be necessary. Two of those, that those competencies are "important for the job market" and "a necessary skill for science—and the humanities" may be worth considering. From the perspective I am taking in this contribution, his argument that they are "necessary for everyone in a data driven world" is more interesting. Matzner on what is necessary here: Data literacy, media literacy, and privacy literacy, from his point of view, may allow the students to take informed

In Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018): *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts.* Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

decisions regarding their own actions and to judge social, technical, and political developments and agendas.

The reason for integrating data science into school curricula would thus not be to educate the next generation of data scientists, but rather to have it contribute to general education in the sense that the students can be rational actors as well as informed citizens. (Matzner showed that the first aspect comes with two major problems, responsibilization and rational bias; nonetheless, both aspects seem important to me.)

WHICH GOALS ARE TO BE ACHIEVED? WHAT SHOULD STUDENTS BE LEARNING?

These considerations already define the goals that from this perspective are to be achieved when dealing with data science in school. Unfortunately, many of the aspects of data literacy, media literacy, and privacy literacy are beyond the scopes of both mathematics and computer science, resulting in the question what the aims and objectives might be in relation to those subjects. Due to my background, I will concentrate here on the computer science side.

Unfortunately, at this point I am not yet able to define the fundamental ideas behind data science. This, however, would be helpful in defining the goals to be achieved. As said before, data science uses tools and methods that have been around in computer science for a long time, such as data bases and visualization techniques, but also tools and methods that have gained momentum recently, such as machine learning (cf. Schutt & O'Neil 2013). While it may be worth considering teaching each of these subdomains (visualization, data bases, machine learning…) in schools, two questions arise: Is there anything in data science contributing to the three types of literacy (data, media, privacy) that goes beyond those subdomains? And if so, which of the subdomains need to be understood in order to understand data science and which can be treated as black boxes without compromising the goals?

One aspect that was largely absent from the symposium was the fact that data analysis is now in many cases part of the data product—from Amazon recommendations to Apple Music to fitness apps to Google search. In all of these cases, the data analysis is being done in real time, being immediately fed back into the system, influencing the "behavior" of the system and thus the behavior of the user. This kind of application is at the core of computer science in the sense that new methods and tools are being developed here. Also, these applications have significant impact in the real world (Cf. echo chambers and filter bubbles as mentioned by Bettina Berendt (2018)).

Another major aspect that was discussed only briefly is the topic of privacy and de-anonymization (or rather de-pseudonymization). Privacy regulations in Europe have significant impact on data science and that impact will increase with the EU General Data Protection Regulation. While legal matters are probably not so much of interest in mathematics and computer science teaching in schools, the reasons why that regulation exists are right at the core of those subjects. As several publications have shown, the robust anonymization of data large data sets is not only an important problem to solve but also poses interesting questions that can be dealt with within mathematics and computer science (cf. Narayanan & Shmatikov, 2008; or de Montjoye et al., 2013).

CONCLUSION

From my point of view, the goal of introducing data science into school curricula cannot be educating the next generation of data scientists. Rather the question is: What does data science contribute to general education? Why is there a need to teach data science to students at school? Several talks have given good reasons. However, many of those aspects that justify teaching data science in schools are beyond the subjects of mathematics and computer science. This leads to the question whether data science should be taught within those subjects or if, instead, teaching it in co-operation with teachers from all relevant subjects makes more sense. I have tried to give some ideas of what the contribution from computer science might be, yet more questions remain open to me than have been answered. Also, the qualifications needed by teachers to address these topics—either within their own subject or in cooperation with other subjects—are not yet obvious to me.

REFERENCES

Berendt, B., & Dettmar, G. (2018). If you're not paying for it, you are the product: A lesson series on data, profiles, and democracy. In Biehler, R. et al. (eds.). *Paderborn Symposium on Data Science Education 2017: The Collected Extended Abstracts*. Paderborn: University of Paderborn.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3, 1376*, 1–5.

Matzner, T. (2018). Data science education as contribution to media ethics. In Biehler, R. et al. (eds.). *Paderborn Symposium on Data Science Education 2017: The Collected Extended Abstracts*. Paderborn: University of Paderborn.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *IEEE Trans. Secur. Priv. 8*, 111–125.

Schutt, R., & O'Neil, C. (2013). *Doing data science*. Cambridge: O'Reilly.