Preprint from Braschler, Stadelmann, Stockinger (Eds.): *"Applied Data Science - Lessons Learned for the Data-Driven Business"*, Springer , 2018 (to appear).

Lessons Learned from Challenging Data Science Case Studies

by Kurt Stockinger, Martin Braschler, and Thilo Stadelmann.

In this chapter, we revisit the conclusions and lessons learned of the chapters presented in Part II of this book and analyze them systematically. The goal of the chapter is threefold: firstly, it serves as a directory to the individual chapters, allowing readers to identify which chapters to focus on when they are interested either in a certain stage of the knowledge discovery process or in a certain data science method or application area. Secondly, the chapter serves as a digested, systematic summary of data science lessons that are relevant for data science practitioners. And lastly, we reflect on the perceptions of a broader public towards the methods and tools that we covered in this book and dare to give an outlook towards the future developments that will be influenced by them.

1. Introduction

Part II of this book contains 16 chapters on the nuts and bolts of data science, divisible into fundamental contributions, chapters on methods and tools, and texts that apply the latter while having a specific application domain in focus. Some of these chapters report on several case studies. They have been compiled with the goal to stay relevant for the readership beyond the lifetime of the projects underlying the specific case studies. To establish this book as a useful resource for reference in any data science undertaking, this chapter serves as a key to unlock this treasure.

The chapter is organized as follows: Section 2 presents a taxonomy that covers the main dimensions of content in the individual chapters previously presented in Part II. In Section 3, we give concise summaries of all chapters and their learnings. On this basis, we then provide an overall aggregation of the lessons learned in Section 4, together with more general insights. Final conclusions are drawn in Section 5.

2. Taxonomy

Table 1 provides a taxonomy covering the content of the case studies described in Part II. The taxonomy highlights the main items of the individual chapters and serves as a structured index for the reader to navigate Part II.

Taxonomy	Discussed in Chapters															

Main focus	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Fundamentals of Data science	x	x	x													
Methodology or algorithm				x	x	x	x	x	x	x			x			x
Tool							x		x						x	
Application		x	x							x	x	x	x	x	x	x
Survey or tutorial				x	x			x			x					
Stages in knowledge discovery process[1]	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Data recording	x							x	x	x				x	x	x
Data wrangling	x				x			x			x	x			x	
Data analysis	x			x	x	x	x	x	x	x	x	x	x	x		x
Data visualization and/or interpretation	x		x	x			x				x	x				x
Decision making	x		x				x			x			x	x		
Competence area[2]	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Technology								x	x					x	x	x
Analytics				x	x	x			x	x	x	x	x	x		x
Data Management						x	x		x		x	x		x	x	x
Entrepreneurship		x	x					x								
Communication							x					x				
Subdisciplines	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**

[1] See Chapter 2, Section 3.6.
[2] See Chapter 3, Figure 1.

Simulation										x				x		
Data modeling				x							x					
Data warehousing											x			x	x	
Big data technology[3]									x					x		x
Information retrieval						x										
Clustering					x						x	x				
Classification[4]					x			x				x				x
Regression				x						x						x
Anomaly detection					x			x	x							
Forecasting				x									x	x		
Visualization				x			x					x				
Security								x								
Governance															x	
Ethics			x													
Tools	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
IT infrastructure[5]	x					x		x	x		x			x	x	x
Off-the-shelf analytics solution																
Scripted analytics				x	x	x		x		x	x	x	x			x
Open source software				x	x		x		x	x	x	x		x		x

[3] E.g. parallel processing, stream processing, etc.
[4] E.g. deep neural networks, SVMs, etc.
[5] E.g. databases, middleware, etc.

Commercial software							x					x				
Off-the-shelf visualization solution							x					x				
Scripted visualization				x	x		x				x	x				
Data modalities	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Numerical data	x		x	x	x		x	x	x	x	x	x	x	x		
Text						x	x	x			x	x				
Images	x				x											x
Audio[6]					x											
Time series	x			x	x		x					x				
Transactional data								x			x	x	x			
Open data							x									
Application domain	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Research	x			x	x	x	x		x						x	
Business			x			x	x		x	x				x		
Biology					x											x
Health	x			x			x					x			x	x
eCommerce and retail			x					x			x		x			
Finance			x											x		
IT								x								

[6] E.g. speech, music

Industry and manufacturing					x					x						
Services	x		x		x						x					

3. Concise reference of individual lessons learned

In this section, we provide a reference to the distilled lessons learned of each chapter of Part II. The section can thus serve the reader to assess their level of data science knowledge and pick out the most pertinent areas for further study.

Chapter 8: What is Data Science?
A treatise of the fundamentals of data science and data science research from a senior researcher's perspective.

Lessons learned:

- Data science is an emerging paradigm for accelerated discovery in any field of human endeavor based on the automated analyses of all possible correlations. It has no tools to establish causality between the observed relationships.
- Maturity of data science as a discipline is approximately a decade ahead and will depend on (a) general principles applicable equally to all domains; and (b) collaboration of experts across previous disciplinary silos (which needs a "chief scientific officer" role).
- Based on the analysis of 150 use cases, a generic 10-step data science workflow (in extension of the knowledge discovery process from Chapter 2) is presented and exemplified based on three major scientific projects.

Chapter 9: On Developing Data Science
Suggests the 20th century hardware-software virtuous innovation cycle as a role model for how data science projects and the discipline itself should be furthered.

Lessons learned:

- Data science is inherently an applied science that needs to be connected to real-world use cases: "necessity is the mother of invention", and data scientists even in research profit from solving pressing problems of businesses.
- Still, data science is more than doing data science projects, and data science research units need to be more than the sum of their parts, contributing to data science "per se" by developing software platforms and generally applicable methodology across domains.
- Several common misunderstandings regarding the adoption of data science in businesses are addressed, including "data science is expensive" or "it is all about AI".

Chapter 10: The Ethics of Big Data Applications in the Consumer Sector
An introduction to and guidelines for ethical considerations in data science applications is given, helping with questions like "to whom does the data belong", or "how is (and should) autonomy, privacy and solidarity (be) affected".

Lessons learned:

- A practical guideline regarding unwanted ethical effects is this: would customers still use the product or provide the data if they knew what their data is used for? What could incentivize them to continue doing it if they knew?
- Trust and acceptance of data science applications can be created by informing the customers transparently, and by always providing an option to chose.
- Based on 5 case studies, a practical weighing of the core values of autonomy, equality, fairness, freedom, privacy, property-rights, solidarity and transparency that can be adopted in a cookbook fashion.

Chapter 11: Statistical Modelling
A plea for the use of relatively simple, traditional statistical modelling methods (also in contrast to "modern black box approaches"). How to maximize insight into model mechanics, and how to account for human interventions in the modelling process.

Lessons learned:

- Descriptive analysis requires explicit statistical models. This includes concrete knowledge of the model formulation, variable transformations and the error structure.
- Statistical models can and should be verified: check if the fit is in line with the model requirements and the subject matter knowledge.
- To obtain sound results and reliable interpretations, the data-generating mechanism within the model developing process and during model assessment have to be considered.

Chapter 12: Beyond ImageNet - Deep Learning in Industrial Practice
An introduction to various case studies on deep learning beyond classifying images: segmentation, clustering, anomaly detection on documents, audio and vibration sensor signals.

Lessons learned:

- For designing a deep neural network, start with a simple architecture and increase the complexity when more insights into the data and model performance are gained. Generally, if a human expert sees the pattern in the data, a deep net can learn it, too.
- There are many options to deal with limited resources, especially limited training data: transfer learning, data augmentation, adaptable model architectures, or semi-supervised learning. Applying deep learning does not need gigabytes of data.
- Deep models are complex, but far from being black boxes: in order to understand the model performance and the learning process, "debugging" methods such as visualizing the learned weights or inspecting loss values are very helpful.

Chapter 13: The Beauty of Small Data - An Information Retrieval Perspective
Discussion and case studies that show the different challenges between leveraging small and big data.

Lessons learned:

- Finding patterns in small data is often more difficult than in big data due to the lack of data redundancy.
- Use stemming to increase the occurences of terms in small document collections and hence increase the potential redundancy to find patterns.

- Enrich data with additional information from external resources and synthesize new, additional keywords for query processing based on relevance feedback.

Chapter 14: Narrative Information Visualization of Open Data
Overview of open data portals of the US, the EU and Switzerland. Description of visualization applications on top of open data that enable narrative visualization: a new form of Web-based, interactive visualization.

Lessons learned:

- Data preparation: The most time consuming aspect of information visualization. Data needs to be manually transformed, harmonized, cleaned and brought into a common data model that allows easy visualization.
- Visualization technology: High level visualization frameworks that enable quick prototyping often cannot be used out of the box. In order to get full visualization flexibility, interactive information visualization and especially narrative visualization often require a development path from rapid prototyping using "out-of-the-box" data graphics towards "customized" visualizations that require some design and coding efforts.

Chapter 15: Security of Data Science and Data Science for Security
A survey on the aspect of computer security in data science (vulnerability of data science methods to attacks; attacks enabled by data science), and on the use of data science for computer security.

Lessons learned:

- Protect your information systems with suitable security controls by rigorously changing the standard privacy configurations, and using a secure software development life cycle (SSDLC) for all own developments.
- Guidelines are given in the "CIS top twenty security controls", and current security issues are posted e.g. in the "OWASP top 10" for web applications.
- Also secure your models: anonymization is not perfect, analysis on encrypted or anonymized data is still under research, and attackers might try to exploit data-driven applications by data poisoning, model extraction etc.

Chapter 16: Online Anomaly Detection over Big Data Streams
Various anomaly detection strategies for processing streams of data in an Apache Spark big data architecture.

Lessons learned:

- Make sure that data processing is performed efficiently since data can be lost in case the stream processing buffers fill up.
- Pearson correlation and event counting work well for detecting anomalies with abrupt data changes. For detecting anomalies based on gradually occurring changes, use relative entropy measures.
- Use resampling techniques to determine statistical significance of the anomaly measure. When annotated ground truth data is available, use supervised machine learning techniques to automatically predict the anomaly type.

Chapter 17: Unsupervised Learning and Simulation for Complexity Management in Business Operations

A study on developing a purely data-driven complexity measure for industrial products in order to reduce unnecessary drivers of complexity, made difficult by the unavailability of data.

Lessons learned:

- In cases where low-level data is unavailable, available high-level data can be turned into a simulation model that produces finer-grained synthetic data in arbitrary quantity, which in turn can be used to train a machine-learning model with the ability to generalize beyond the simulation's discontinuities.
- Complexity of industrial product architectures and process topologies can be measured based on the minimum dimensionality of the bottleneck layer of a suitably trained autoencoder.
- Data-driven complexity measurement can be an alternative to highly qualified business consultants, measuring complexity in a fundamentally different but result-wise comparable way.

Chapter 18: Data Warehousing and Exploratory Analysis for Market Monitoring

An introduction to data warehouse design, exemplified by a case study for an end-to-end design and implementation of a data warehouse and clustering-based data analysis for e-commerce data.

Lesson learned:

- Data warehouse design and implementation easily take 80% of the time in a combined data preparation and analysis project, as efficiently managing a database with dozens of tables of more than 10^7 records requires careful database tuning and query optimization.
- Data from anonymous e-commerce users can be enriched using Google Analytics as a source; however, the data quality of this source is not easily accessible, making results based on this source to be best considered as estimates.
- When using clustering as an instance of unsupervised machine learning, the necessary human analysis of the results due to the unavailability of labels can be eased using sampling: verify a clustering by analyzing some well-known clusters manually in detail.

Chapter 19: Mining Person-Centric Datasets for Insight, Prediction, and Public Health Planning

A data mining case study demonstrating how latent geographical movement patterns can be extracted from mobile phone call records, turned into population models, and utilized for computational epidemiology.

Lessons learned:

- Data processing for millions of individuals and billions of records require parallel processing toolkits (e.g., Spark); still, the data needed to be stored and processed in aggregated form at the expense of more difficult and expressive analysis.
- It is important to select the right clustering algorithm for the task (e.g., DBSCAN for a task where clusters are expressed in different densities of the data points, and K-means where clusters are defined by distances), and to deal with noise in the measurements.

- Visualization plays a major role in data analysis: to validate code, methods, results; to generate models; and to find and leverage to wealth of unexpected, latent information and patterns in human-centric datasets.

Chapter 20: Economic Measures of Forecast Accuracy for Demand Planning - A Case-Based Discussion

Methods for evaluating the forecast accuracy to estimate the demand of food products.

Lessons learned:

- Error metrics are used to evaluate and compare the performance of different forecasting models. However, common error metrics such as root mean square error or relative mean absolute error can lead to bad model decisions for demand forecasting.
- The choice of the best forecasting model depends on the ratio of oversupply costs and stock-out costs. In particular, a baseline model should be prefered over a peak model if the oversupply costs are much higher than the stock-out costs and vice versa.
- Choosing the optimal observation time window is key for good quality forecasts. A too small observation window results in random deviations without yielding significant insights. A too large observation window might cause poor performance of short term forecasts.

Chapter 21: Large-Scale Data-Driven Financial Risk Assessment

Study of an approach to standardize the modeling of financial contracts in view of financial analysis, discussing the scalability using Big Data technologies on real data.

Lessons learned:

- Computational resources nowadays allow solutions in finance, and in particular in financial risk analysis, that can be based on the finest level of granularity possible. Analytical shortcuts that operate on higher levels of granularity are no longer necessary.
- Financial (risk) analysis is possible at the contract level. The analysis can be parallelized and distributed among multiple computing units, showing linear scalability.
- Modern Big Data technologies allow the storage of the entire raw data, without pre-filtering. Thus, special purpose analytical results can be created quickly on demand (with linear computational complexity).
- Frequent risk assessment of financial institutions and ultimately the whole financial system is finally possible on a level potentially on par with that of other fields such as modern weather forecasts.

Chapter 22: Governance and IT Architecture

Governance model and IT architecture for sharing personalized health data.

Lessons learned:

- Citizens are willing to contribute their health data for scientific analysis if they or family members are affected by diseases.
- Data platforms that manage health data need to have highly transparent governance structures, strong data security standards, data fusion and natural language processing technologies.

- Citizens need to be able to decide by themselves for which purpose and with whom they share their data.

Chapter 23: Image Analysis at Scale for Finding the Links between Structure and Biology
End to end image analysis based on big data technology to better understand bone fractures.

Lessons learned:

- Image data are well-suited for qualitative analysis but require significant processing to be used in quantitative studies.
- Domain specific quantitative metrics such as average bone thickness, cell count or cellular density need to be extracted from images before they can be correlated to images and other data modilites.
- Rather than removing data samples with missing values, data quality issues can be handled by imputation, bootstrapping and incorporating known distributions.

4. Aggregated insights

On the basis of the individual lessons learned that we described in the previous section, we will now provide an overall condensation of the lessons learned. We feel that these points are highly relevant and that they form a concise set of "best practices" that can gainfully be referenced in almost every data science project.

- Data science is an inherently interdisciplinary endeavour and needs close collaboration between academia and business. To be successful in a wide range of domains, close *collaboration and knowledge exchange* between domain experts and data scientists with various backgrounds is essential.
- Building a *trust relationship* with customers early on by providing transparent information about the data usage along with rigorous data security practices is key to guarantee wide adoption of data products. Let the customers choose which data they want to share with whom. Part of building trust is also to care for potential *security* issues in and through data analysis right from the start.
- *Data wrangling,* which includes transforming, harmonizing and cleaning data, is not only a vital prerequisite for machine learning but also for visualization and should thus be a key effort of each data science project. Ideally, data wrangling should be automated using machine learning techniques to ease the burden of manual data preparation.
- Leverage existing *stream processing frameworks* for enabling data wrangling and analysis in real time.
- When choosing a machine learning model to solve a specific problem, start with *simple algorithms* where only a small number of hyperparameters need to be tuned and a simple model results. Increase the complexity of the algorithms and models if necessary and as more insights into the data and model performance are gained.
- Use *visualization* to gain insights into data, track data quality issues, convey results, and even understand the behavior of machine learning models (see also below).
- Modern *big data technology* allows storing, processing and analyzing vast amounts of (raw) data – often with linear scalability. Restricting models to representative data samples for the sake of reducing data volumes is not strictly necessary any more.

- Leveraging *small data* with low redundancy requires different and maybe more sophisticated approaches than leveraging *big data* with high redundancy.

In condensing the lessons learnt to best practices that are generalizable, there is a danger of losing the surprising, inspiring insights that only more detailed looks at specific contexts can bring. By necessity, it is impossible to exhaustively compile such "inspiration" in a list. However, we very much think that much of this inspiration can be found between the covers of this book. In reflecting on the journey of the book's creation, on our own experiences with data science projects over the years, and on the collaboration with the excellent colleagues that have contributed to this volume, we want to emphasize some of these "highlights" that we found:

Data science education has to be interdisciplinary and above Bachelor level to ensure the necessary skills also for societal integration. What are useful outcome competencies for data scientists? The answer to this question differs for data scientists focusing on the engineering aspect compared to those specializing on business aspects or communication or any application domain. But they all will have the following in common: an understanding of the core aspects and prospects of the main methods (e.g., machine learning), tools (e.g., stream processing systems) and domains (e.g, statistics) as well as experience in hands-on projects (in whatever role in an interdisciplinary team). This, combined with the maturity that comes with completed discipline-specific studies during one's Bachelor years, enables a data scientist to ponder and weigh the societal aspects of work in a responsible and educated manner.

Data-driven innovation is becoming increasingly fast, yet not all innovation is research-based; that is why networks of experts are becoming more important to find the right ideas and skills for any planned project. In the area of pattern recognition for example, we see a usual turnover time from published research result at a scientific conference to application in an industrial context of about three months. Many of the results there are driven by deep learning technology, and the lines between fundamental and applied research have become reasonably blurred in recent years (with companies producing lots of fundamental results, and universities engaging in many different application areas, compare e.g. Stadelmann et al. (2018)). This speaks strongly for collaborations between scientists and engineers from different organisations and units that complement each other's knowledge and skills, e.g. from the academic and industrial sector. Simultaneity in working on the fundamental aspects of methods (e.g. furthering deep learning per se) and making it work for a given problem by skillful engineering (for example by clever problem-dependent data augmentation and a scalable hardware setup) seems to be key.

On the other hand, only one third of data-driven innovation needs novel research to happen in order to take place - two thirds are implementable based on existing technology and tools once the party in need of the innovation gets to know the availability or feasibility of the endeavour, given that resources are available (Swiss Alliance for Data-Intensive Services, 2018). If two thirds of the innovation potential in a country like Switzerland are achievable by education (informing stakeholders about possibilities) and consulting (bringing in expert knowledge on how to approach the sought innovation), this is a strong argument for every interested party to team

up with like-minded organizations and individuals, again to complement each other's skills and know-how to *"together move faster"*[7].

The paradigm of data parallelism that is enabled by state of the art big data technology makes designing parallel programs relatively easy. However, fully understanding their performance remains hard. Writing scalable, parallel or distributed programs has generally been considered hard, especially when data is not read-only but can be updated. The main challenge is how to solve the "critical section" (Quinn 2003), i.e. how to avoid that two program threads update a specific data item at the same time and thus result in data inconsistency. Different communities use different approaches to tackle this problem. One of the lowest level concepts for parallel programming is to use multithreading, which requires explicit handling of the "critical section" via semaphores (Kleiman, Shah et al., 1996). The high-performance community typically uses a higher level of abstraction based on "message passing" where parallel processes communicate via explicit messages (Gropp, Gropp et al., 1999). Both approaches require highly skilled people to write efficient programs that scale and do not result in deadlocks. The paradigm of data parallelism deployed by state-of-the-art big data technology such as Apache Spark enables implicit parallelism (Zaharia, Xin, et al., 2016). By design, the core data structures such as Resilient Distributed Datasets or Dataframes enable parallel processing based on the MapReduce paradigm where the programmer has only little design choices to influence the program execution.

This implicit parallelism has the great advantage that even people without deep knowledge of parallel programming can write programs that scale well over tens or hundreds of compute nodes. However, the implicit parallelism also comes with a big disadvantage – namely the illusion that programs scale "by default" and that "parallel programming becomes easy". The hard part of writing good parallel programs with novel big data technology is to fully understand the complex software stack of a distributed system, the various levels of distributed memory management and the impact of data distribution on the runtime of SQL queries or machine learning algorithms. Hence, detailed performance analyses of the workloads and manual optimization techniques such as task re-partitioning based on workload characteristics is often the best solution to overcome potential performance problems. The important takeaway message is that understanding and tuning the performance of big data applications can easily take a factor of 10 more time than writing a program that leverages big data technology.

Let machine learning and simulation complement each other. The traditional scientific approach is often based on experimentation and simulation (Winsberg, 2010). Experiments are carefully designed based on a specific model. Once data is available or produced by (physical) experiments, the certain phenomena of interest can be evaluated empirically. In addition, simulation is used to complement experimentation. Hence, simulation can be used to verify experiments, and experiments can be used to adapt the simulation model. By comparing experimental outcomes with those from simulation, the degree of current understanding of the observed phenomenon (as encoded in the simulation) can be assessed. However, the disadvantage of this approach is that building experiments can be very time consuming and costly. For instance, building a high-energy physics experiment end-to-end can take more than 10 years (Brumfiel 2011). Moreover, there might not be enough data available to run statistically

[7] See https://data-service-alliance.ch/ for an example of implementing this principle in a way the three authors of this chapter are involved in.

significant experiments. Finally, building simulation models might become extremely complex, in particular, when some physical, chemical or biological processes are not fully understood yet.

Hence, machine learning can be applied as an additional pillar. In traditional experimental science, machine learning can be used to *learn* a model from both the experimental and simulated data. The resulting model has the potential to generalize beyond the discontinuities of the simulation model, thus relieving one from making the simulation overly complex. This is not to replace experimentation and simulation, but in addition. On the other hand, in other fields of data science, simulation can serve as a means to data synthesis, thus enhancing the available training data for machine learning approaches. This is heavily used under the umbrella term of "data augmentation" for example in the field of deep learning.

Models learned from data need to be robust and interpretable to facilitate "debugging" and make them acceptable to humans. Statistical or machine learning models are usually subject to a comprehensive empirical evaluation prior to deployment; the results of these experiments have the power to both show the respective strengths and weaknesses of the model as well as to demonstrate their reliability and generalization capabilities to a critical reviewer (e.g., a business owner, customer, or human subject to a machine-supported decision). Yet, we as humans feel generally uncomfortable when we are subject to processes that we cannot fully grasp and at which's mercy we feel we are (Lipton, 2018); and as developers, having no insight into complex processes like machine learning pipelines and training processes hinders debugging and effective optimization of the model (Stadelmann et al., 2010).

Recent research and development into model interpretability (see e.g. (Ng, 2016), (Shwartz-Ziv & Tishby, 2017), or (Olah et al., 2017)) not only allows the statement that even the most seemingly opaque machine learning models like deep neural networks can be comprehended to a large degree by humans. The respective work also opens up many more possible developments in research (through a better understanding of what goes wrong) and specific high-risk application domains like automated driving or clinical health (due to the ability to fulfill regulations and bring about necessary performance gains). Thus, trust can be built in applications that directly face a human customer; and better understanding by developers also brings about more robust models with less peculiar behaviour (compare (Szegedy et al., 2013) with (Amirian et al., 2018)). Moreover, the understanding possible through introspection into models enables data scientists that are mere users of machine learning to select the best fitting approach to model the data at hand - a task that otherwise needs intimate knowledge of the inductive biases (Mitchell, 1997, ch. 2) of many potential methods as well as of the structure of the given data.

5. Conclusions

Data science is a highly interesting endeavour, breaking new ground in many ways. Due to the young age and the wide range of the discipline, a number of myths have already taken deep hold, most prominently those that lead to exasperated outbursts along the lines of "no one knows how these algorithms work" or "no one can understand why the output looks like this". We claim that this is plainly untrue, and the various case studies covered in Part II of this book are an excellent testament to this: there is a wide range of scientific literature, and an

abundance of tools and methods available to data science practitioners today; there is a wealth of well-founded best practices on how to use them, and there are numerous lessons learned waiting to be studied and heeded.

5.1 Deconstructing myths by the example of recommender services

If we look at the disruptive players in the information space and their platforms, such as Facebook, Google, Amazon and others, they also very much rely on these tools and methods to drive their services. Many of the phenomena that e.g. recommender services exhibit in their selection of items are indeed fairly easily and conclusively interpretable by those that have studied the relevant, well-documented algorithms.

It follows that discussions about whether such machine learning components exhibit unwanted biases are certainly very pertinent, but oftentimes not led in the most effective manner (see e.g. the discussion on biases in word embeddings by Bolukbasi et al., (2016)). The rapidly increasing use of recommenders based on machine learning to support many knowledge-intensive processes such as media consumption, hiring, shopping etc. is observed with anxiety by some of those that used to enjoy influence in these fields. Unfortunately, however, these discussions on the merits of machine-generated recommendations are many times led under the wrong pretext. Often the starting point is whether the operators of the recommender service follow a sinister agenda, for example, feeding consumers a steady diet of questionable information of very little variety ("filter bubble", see Pariser (2011)). In this view, compounding the sinister agenda of the operator is, again, the fact that "nobody knows what they are doing and how they do it". Scenarios such as "artificial intelligence is already making hiring decisions and your every blink is going to influence your chances" are talked up.

Looking at the situation more soberly, and abstracting from the source of a decision - be it human or machine - the question should be: what do we really want as the output? And does a human (as the chief alternative to the AI-based recommender system) deliver it better and with less bias? In a sense, algorithms can exhibit traits that are very human: if the data used for training exhibits unwanted biases, so will the output of the recommender. A widely reported instance of this was the Microsoft chatbot "Tay" that quickly learned abusive and racist language from Twitter feeds (Hunt, 2016).

Reflecting on the filter bubble, the narrow focus of the information stream delivered to some consumers can easily be an expression of overfitting - of the hard problem to generalize to things unseen in prior training, and in incorporating aspects beyond mere item similarity, such as novelty, diversity etc. into the selection mechanism.

Which closes the circle and brings us back to the all-important question: what do we want from our data? Do we want a "superhuman result" - insight that a human could not have gleaned from the data, or behaviour that a human would not exhibit? Or do we want to emulate the (competent) human, producing the same decision a human expert would have arrived at, potentially faster or at lower cost? Are we open to new insights, and can machine-generated recommendations augment human decision-making by delivering complementary information, being able to leverage (volumes of) information that humans cannot process? Can it even help to overcome human bias?

5.2 Outlook to a data-driven society

In an abstract perspective, a recommendation - be it made by a human or a computer - is the output of a function of the case-specific inputs plus a number of parameters inherent to the instance making the recommendation, such as preferences and previous history. Two human experts will produce different recommendations given the same inputs. Analogously, the output of an algorithm will change as we change the parametrization. Human decision makers are often bound by rules and regulations in their freedom to make decisions. In the course of the evolution of civilization, there has been constant debate on how to shape these rules and regulations, whom to grant the power to define them, and who to task with enforcing them. Unsurprisingly, we are not at the end of this road. We see no fundamental reason why similar rules and regulations cannot influence the parametrization, and thus the operation of for example recommender services.

Data science in general has not only the ability to automate or support decision processes previously reserved to capable humans only, at scale; it also has the potential to alter the ways our societies work in disruptive ways. Brooks (2017) skillfully disarms unsubstantiated fears of mass unemployment in the next decade, and multitudes of humanoid robots or the rise of human-level artificial intelligence are nowhere to be seen. But the current technological possibilities paired with contemporary economic incentives make it quite clear that society will be impacted on a fundamental level: how can this debate be held in a constructive way in the face of the opinion economy on social media? How to distribute work when repetitive jobs (e.g., medical diagnose, legal case research, or university-level teaching) get digitized to some degree? How to fill one's time in a meaningful way and distribute the gain from increased economic efficiency fairly if it is generated by algorithms in large corporations?

With these exemplary questions above we do not foremost promote to engage in research on "data science for the common good" (see e.g. (Emrouznejad & Charles, 2018)), although this is important. We rather suggest that much more than thinking about rules of how humans and technology can get along and interact in the future, the possibilities presented to us through a wider deployment of data science will bring us to deal with an age-old topic: how do we want to get along with our fellow human beings. It is a question of society, not technology, to decide on how we share the time and other resources made available to us through the value generated from data. Whom we let participate (education), profit (economy) and decide (politics). A big challenge lies ahead in having such a meaningful dialog between technological innovators (and chiefly among them, data scientists), and stakeholders from government and society.

As it is hard not only to predict, but also to imagine a future that deviates largely from a simple extrapolation of today, it is very helpful to recall some of the scenarios that researchers and thinkers have created. Not because they are necessarily likely or desirable, but because seeing a vivid mental picture of them could help in deciding if these scenarios are what we want - and then take respective action. There is Kurzweil's (2010) vision of a superhuman artificial intelligence that controls everything top-down. It can be contrasted with the bottom-up scenario of digitally enabled self-organization suggested by Helbing (2015) that is based on today's technology. Pearl and Mackenzie's (2018) observe as well that current artificial intelligence is limited as long as it cannot use causation (and thus cannot imagine new scenarios), thus outruling superintelligence in the medium term. Harari (2016) puts future influences of massively

applied data science on the job market in the center, exploring the possibilities of how humans augment (instead of supersede) themselves with biotechnology, robotics and AI, but creating a new class of unemployables. Future "class" differences are also a major outcome of the data-driven analyses of Piketty (2014). Precht's (2018) utopia finally reestablishes the humanitarian ideal of working just to better ourselves and the rest of humanity, funded by the profit generated by increasing automatization. We encourage the reader to dive into the original sources of these heavily abbreviated scenario descriptions to see potential consequences of today's developments in pure (and thus often extreme, thus unrealistic) form.

In the end, these sophisticated scenarios may suggest the following prime challenges of society when dealing with the opportunities and risks of data science applied largely and at scale: the "shaping of the future" is not a technical-scientific undertaking, but takes larger efforts (foremost politically, to change regulatory frameworks that still work but are unfit for changed circumstances as are likely to happen). Change could be driven by a societal consensus on how collaboration in the future should function (the digital technology works as a means to this collaboration), when we overcome the urge to let short-time gains in convenience take us down a path of advancement to an unimagined end. Opportunities, both for individual stakeholders in businesses and industry as well as for societies, are large. Risks exist, mitigations likewise. We suggest to take the lessons learned so far, some of them collected in this volume, and create places - at work, at home, on earth - worthy of living in and working for.

References

Amirian, M., Schwenker, F., & Stadelmann, T. (2018). Trace and detect adversarial attacks on CNNs using feature response maps. In: *Proceedings of the 8th IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition* (ANNPR), Siena, Italy, September 19–21, 2018. IAPR.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357).

Brooks, R. (2017). The Seven Deadly Sins of AI Predictions. *MIT Technology Review*. Available online (March 28, 2018): https://www.technologyreview.com/s/609048/the-seven-deadly-sins-of-ai-predictions/.

Brumfiel, G. (2011). High-energy physics: Down the petabyte highway. *Nature News*, *469*(7330), 282-283.

Emrouznejad, A., & Charles, V. (2018). *Big Data for the Greater Good*. Springer.

Gropp, W. D., Gropp, W., Lusk, E., & Skjellum, A. (1999). *Using MPI: portable parallel programming with the message-passing interface* (Vol. 1). MIT press.

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Random House.

Helbing, D. (2015). *Thinking ahead-essays on big data, digital revolution, and participatory market society*. Springer.

Hunt, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. The Guardian, 24.

Kleiman, S., Shah, D., & Smaalders, B. (1996). *Programming with threads* (p. 48). Mountain View: Sun Soft Press.
Kurzweil, R. (2010). The singularity is near. Gerald Duckworth & Co.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue*, 16(3), 30, pp. 31-57, ACM.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Ng, A. (2016). Nuts and bolts of building AI applications using Deep Learning. *NIPS* Keynote talk, available online (July 26, 2018): https://media.nips.cc/Conferences/2016/Slides/6203-Slides.pdf. .

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, 2(11), e7. Available online (July 26, 2018): https://distill.pub/2017/feature-visualization/.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Piketty, T. (2014). *Capital in the 21st Century*. Harvard University Press.

Precht, R. D. (2018). Hunters, Herdsmen, Critics. A utopia for digital society. Goldmann.

Quinn, M. J. (2003). Parallel Programming. *TMH CSE*, *526*.

Shwartz-Ziv, R., & Tishby, N. (2017). *Opening the black box of deep neural networks via information*. arXiv preprint arXiv:1703.00810.

Stadelmann, T., Amirian, M., Arabaci, I., Arnold, M., Duivesteijn, F. F., Elezi, I., Geiger, M., Lörwald, S. Meier, B. B., Rombach, K., & Tuggener, L. (2018). Deep Learning in the Wild. In: *Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition* (ANNPR'18), Siena, Italy, September 19-21.

Stadelmann, T., Wang, Y., Smith, M., Ewerth, R., & Freisleben, B. (2010). Rethinking Algorithm Design and Development in Speech Processing. In: *Proceedings of the 20th IAPR International Conference on Pattern Recognition* (ICPR'10), Istanbul, Turkey, August 23-26.

Swiss Alliance for Data-Intensive Services (2018). *Digitization & Innovation through cooperation. Glimpses from the Digitization & Innovation Workshop at "Konferenz Digitale Schweiz"*. Blog post, January 16, 2018, available online (July 26, 2018): https://www.data-service-alliance.ch/blog/blog/digitization-innovation-through-cooperation-glimpses-from-the-digitization-innovation-workshop.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks*. arXiv preprint arXiv:1312.6199.

Winsberg, E. (2010). Science in the age of computer simulation. University of Chicago Press.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Ghodsi, A. (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM*, *59*(11), 56-65.