Report from Dagstuhl Seminar 18401

Automating Data Science

Edited by

Tijl De Bie[1], Luc De Raedt[2], Holger H. Hoos[3], and Padhraic Smyth[4]

1 **Ghent University, BE,** `tijl.debie@ugent.be`
2 **KU Leuven, BE,** `luc.deraedt@cs.kuleuven.be`
3 **Leiden University, NL,** `hh@liacs.nl`
4 **University of California – Irvine, US,** `smyth@ics.uci.edu`

Abstract

Data science is concerned with the extraction of knowledge and insight, and ultimately societal or economic value, from data. It complements traditional statistics in that its object is data as it presents itself in the wild (often complex and heterogeneous, noisy, loosely structured, biased, etc.), rather than well-structured data sampled in carefully designed studies. It also has a strong computer science focus, and is related to popular areas such as big data, machine learning, data mining and knowledge discovery.

Data science is becoming increasingly important with the abundance of big data, while the number of skilled data scientists is lagging. This has raised the question as to whether it is possible to automate data science in several contexts. First, from an artificial intelligence perspective, it is interesting to investigate whether (data) science (or portions of it) can be automated, as it is an activity currently requiring high levels of human expertise. Second, the field of machine learning has a long-standing interest in applying machine learning at the meta-level, in order to obtain better machine learning algorithms, yielding recent successes in automated parameter tuning, algorithm configuration and algorithm selection. Third, there is an interest in automating not only the model building process itself (cf. the Automated Statistician) but also in automating the preprocessing steps (data wrangling).

This Dagstuhl seminar brought together researchers from all areas concerned with data science in order to study whether, to what extent, and how data science can be automated.

Seminar September 30–October 5, 2018 – http://www.dagstuhl.de/18401
2012 ACM Subject Classification Computing methodologies → Artificial intelligence, Computing methodologies → Machine learning
Keywords and phrases artificial intelligence, automated machine learning, automated scientific discovery, data science, inductive programming
Digital Object Identifier 10.4230/DagRep.8.9.154
Edited in cooperation with Jefrey Lijffijt

Automating Data Science, *Dagstuhl Reports*, Vol. 8, Issue 09, pp. 154–181
Editors: Tijl De Bie, Luc De Raedt, Holger H. Hoos, and Padhraic Smyth
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Tijl De Bie (Ghent University, BE)
Luc De Raedt (KU Leuven, BE)
Holger H. Hoos (Leiden University, NL)
Padhraic Smyth (University of California – Irvine, US)

Introduction

Data science is concerned with the extraction of knowledge and insight, and ultimately societal or economic value, from data. It complements traditional statistics in that its object is data as it presents itself *in the wild* (often complex and heterogeneous, noisy, loosely structured, biased, etc.), rather than data well-structured data sampled in carefully designed studies.

Such 'Big Data' is increasingly abundant, while the number of skilled data scientists is lagging. This has raised the question as to whether it is possible to automate data science in several contexts. First, from an artificial intelligence perspective, it is related to the issue of "robot scientists", which are concerned with the automation of scientific processes and which have so far largely focused on the life sciences. It is interesting to investigate whether principles of robot scientists can be applied to data science.

Second, there exist many results in the machine learning community, which has since the early 1980s been applying machine learning at a meta-level, in order to learn which machine learning algorithms, variants and (hyper-)parameter settings should be used on which types of data sets.

In recent years, there have been breakthroughs in this domain, and there now exist effective systems (such as Auto-WEKA and auto-sklearn) that automatically select machine learning methods and configure their hyperparameters in order to maximize the predictive performance on particular datasets.

Third, there are projects such as the Automated Statistician that want to fully automate the process of statistical modeling. Such systems could dramatically simplify *scientific data modeling* tasks, empowering scientists from data-rich scientific disciplines such as bioinformatics, climate data analysis, computational social science, and so on. To ensure success, important challenges not only from a purely modelling perspective, but also in terms of interpretability and the human-computer interface, need to be tackled. For example, the input to the Automated Statistician is a dataset, and the system produces not only a complex statistical model by means of a search process, but also explains it in natural language.

Fourth, there is an interest in not only automating the model building step in data science, but also various steps that precede it. It is well known in data science that 80% of the effort goes into preprocessing the data, putting it in the right format, and selecting the right features, whereas the model-building step typically only takes 20% of the effort. This has motivated researchers to focus on automated techniques for data wrangling, which is precisely concerned with transforming the given dataset into a format that can be handled by the data analysis component. Here, there are strong connections with inductive programming techniques.

Fifth, as it is often easier for non-expert users to interpret and understand visualisations of data rather than statistical models, work on automatic visualisation of data sets is also very relevant to this Dagstuhl seminar.

Finally, an interesting and challenging research question is whether it is possible to develop an integrated solution that tackles all these issues (as is the topic of the ERC AdG SYNTH).

Overview of the seminar

Structure of the seminar

The seminar was structured as follows. The mornings were generally dedicated to presentations (short tutorials on day one), whereas the afternoons were generally dedicated to discussions such as plenary discussions, smaller-group breakout sessions, and flex time that was kept open prior to the seminar. The flex time ended up being dedicated to a mix of presentations and breakout sessions.

Challenges in automating data science

On day one, a range of challenges for research on automating data science were identified, which can be clustered around the following six themes:

1. **Automating Machine Learning (AutoML)**
   Main challenges: computational efficiency; ensuring generalization also for small data; make AutoML faster and more data-efficient using meta-learning; extending ideas from AutoML to exploratory analysis / unsupervised learning.
2. **Exploratory data analysis and visualization**
   Main challenges: the fact that there is there is no single or clearly defined objective; help the user make progress towards an ill-defined goal; (subjective) interestingness of an analysis, a pattern, or a visualization; integrate machine learning and interaction in exploration; exploration of data types beyond simply tabular; veracity of visualizations; how to quantify progress and measure success; the need for benchmarks.
3. **Data wrangling**
   Main challenges: extend the scope of AutoML to include data wrangling tasks; user interfaces to provide intuitive input in data wrangling tasks; how to quantify progress and measure success; the need for benchmarks.
4. **Automation and human-centric data science (explainability, privacy, fairness, trust, interaction)**
   Main challenges: build-in privacy and fairness constraints in automatic data science systems; the dangers of ignorant usage of automated data science systems; different levels of expertise benefit from different degrees of automation; optimizing the performance of the combined human/machine ‘team’; determine when and where the human must be involved; definition or criteria for explainability; risk that automation will reduce explainability and transparency; explainability to whom – a data scientist or layperson?
5. **Facilitating data science by novel querying and programming paradigms**
   Main challenges: interactive data models to help users gain intuitive understanding; declarative approaches for data analysis, querying, and visualization; a query language for automated data science.
6. **Evaluation**
   Main challenges: robust objective measures for data science processes beyond predictive modelling; subjective measures: measures that depend on the user background and goals; evaluation of the entire data science pipeline versus individual steps; reproducibility in the presence of user interactions.

Topics discussed in depth

These identified challenges were then used to determine the program of the rest of the seminar. Talks were held on partial solutions to a range of these challenges. In addition, breakout discussions were held on the following topics:

1. The relation between data-driven techniques and knowledge-based reasoning.
2. Data wrangling.
3. Beyond the black-box: explainability.
4. Automation of exploratory / unsupervised data science tasks, and visualization.
5. Automating data science for human users.

Along with abstracts of the talks, detailed discussions of the main ideas and conclusions of each of these breakout sessions are included in this Dagstuhl report.

Discussion and outlook

Automating data science is an area of research that is understudied as such. AutoML, as a subarea of automating data science, is arguably the first subarea where some remarkable successes have been achieved. This seminar identified the main challenges for the field in translating these successes into advances in other subareas of automating data science, most notably in automating exploratory data analysis, data wrangling and related tasks, integrating data and knowledge-driven approaches, and ultimately the data science process as a whole, from data gathering to the creation of insights and value.

Further developing automated data science raises several challenges. A first challenge concerns the evaluation of automated data science methods. Indeed, the possibility to automate is preconditioned on the availability of criteria to optimize. A second key one is how to ensure that automated data science systems remain Human-Centric, viewing humans as useful allies and ultimate beneficiaries. This can be achieved by designing effective user-interaction techniques, by ensuring explainability, and by ensuring privacy is respected and individuals are treated fairly. These are basic requirements for ensuring justified trust in automated data science systems, and thus key drivers to success.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Automated Machine Learning from Spatio-temporal Data

Mitra Baratchi (Leiden University, NL)

Main reference Mitra Baratchi, Geert Heijenk, Maarten van Steen: "Spaceprint: a Mobility-based Fingerprinting Scheme for Public Spaces", CoRR, Vol. abs/1703.09962, 2017.
URL https://arxiv.org/abs/1703.09962

Spatio-temporal mobility datasets are generated abundantly as a result of prevalent use of location-aware technologies. Incorporating unprecedented information about moving entities such as people, animals, and vehicles, automating the process of learning from such data opens the door towards many applications in ecology, transportation, and urban planning. However, due to having a non-propositional representation, automated machine learning from raw mobility data is still an open challenge. Current machine-learning-based approaches using such data still rely on an extensive manual pre-processing phase. In this talk, I presented two examples of automated pre-processing tasks based on mobility data in the context of space classification and map segmentation. Both these examples can achieve automation through defining an unsupervised learning problem on the original representation of data.

3.2 Towards Automated Clustering

Hendrik Blockeel (KU Leuven, BE)

Joint work of Hendrik Blockeel, Toon Van Craenendonck, Wannes Meert, Sebastijan Dumancic, Elia Van Wolputte
URL https://dtai.cs.kuleuven.be/software/cobras/

This talk provides an overview of the PhD research of Toon Van Craenendonck on semi-automated, interactive clustering. The main conclusions of this research are as follows. The choice of the clustering algorithm strongly affects the results of clustering. Choosing the most suitable algorithm cannot be done with internal quality measures, but it can be done using a small number of must-link and cannot-link constraints. A novel algorithm called COBRAS is proposed that makes use of an intermediate layer between clusters and instances, called super-instances, and that automatically determines the appropriate granularity of the super-instances. COBRAS is the first clustering algorithm that is truly interactive in the sense that it combines three desirable properties: it is anytime, query-efficient, and time-efficient.

References

1 T Van Craenendonck, H Blockeel (2015). Using internal validity measures to compare clustering algorithms. Benelearn 2015 Poster presentations (online), 1-8
2 T Van Craenendonck, H Blockeel (2017). Constraint-based clustering selection. Machine Learning 106(9-10): 1497-1521.
3 T Van Craenendonck, S Dumancic, H Blockeel (2017). COBRA: A Fast and Simple Method for Active Clustering with Pairwise Constraints. IJCAI 2017: 2871-2877

4 T Van Craenendonck, S Dumančić, E Van Wolputte, H Blockeel (2018). COBRAS: Fast, Iterative, Active Clustering with Pairwise Constraints. In Proc. of 17th International Symposium on Intelligent Data Analysis, 2018. Springer, to appear. Preprint arXiv:1803.11060

5 T Van Craenendonck, W Meert, S Dumancic, H Blockeel (2018). COBRAS-TS: A new approach to Semi-Supervised Clustering of Time Series. In Proc. of 20th International Conference on Discovery Science. Springer, to appear. Preprint arXiv:1805.00779

3.3 AutoDiscovery : Intelligent Automated Exploratory Data Analysis for Biomedical Research

Ray G. Butler (Butler Scientifics – Barcelona, ES)

Joint work of Ray G. Butler, Joan Guàrdia-Olmos, Javier Hernández-Losa

According to NIH and SciMago Journal & Country Rank estimations, there are more than 400,000 principal investigators worldwide actively running biomedical research projects in the form of clinical studies, collaborations with pharma companies and basic biological research, among others.

The datasets being produced through these projects are distinguished primarily by their complexity in terms of multidimensionality and sample stratification. John W. Tukey's exploratory data analysis (EDA) techniques are rising in response to this particular scenario. However, both open-source and commercial EDA software packages typically require a broad range of data science skills and knowledge including data integration and visualization, software programming and statistical methodologies which makes it difficult for principal investigators to become actively involved in the exploratory phase.

AutoDiscovery is an intelligent automated exploratory data analysis software that helps biomedical principal investigators integrating and exploring their complex datasets to unveil associations with high statistical significance and clinical relevance hidden in the data files of scientific experiments and clinical trials.

3.4 Elements of an Automated Data Scientist

Luc De Raedt (KU Leuven, BE)

Joint work of Luc De Raedt, Hendrik Blockeel, Samuel Kolb, Stefano Teso, Gust Verbruggen
Main reference Luc De Raedt, Hendrik Blockeel, Samuel Kolb, Stefano Teso, Gust Verbruggen: "Elements of an Automatic Data Scientist", in Proc. of the Advances in Intelligent Data Analysis XVII – 17th International Symposium, IDA 2018, 's-Hertogenbosch, The Netherlands, October 24-26, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 11191, pp. 3–14, Springer, 2018.
URL https://doi.org/10.1007/978-3-030-01768-2_1

We provide a simple but non-trivial setting for automating data science. Given are a set of worksheets in a spreadsheet and the goal is to automatically complete some values. We also outline elements of the SYNTH framework that tackles this task: SYNTH-A-SIZER, an automated data wrangling system for automatically transforming the problem into attribute-value format; TACLE, an inductive constraint learning system for inducing formula's in spreadsheets; MERCS, a versatile predictive learning system; as well as the autocompletion component that integrates these systems.

3.5 Towards a Measurement Theory for Data Science

Peter Flach (University of Bristol, GB)

Performance evaluation is of clear importance in machine learning and data science, and arguably even more so for *automated* data science. Our understanding of performance evaluation measures for machine-learned classifiers has improved considerably over the last twenty years. In this short talk I highlighted a range of areas where understanding is still lagging behind our algorithmic advances, sometimes leading to ill-advised practices in classifier evaluation. I argued that in order to make further progress we need to develop a proper *measurement theory* for data science. I gave some examples what such a measurement theory might look like and what kinds of new results it would entail. In future work I will explore the idea that key properties such as classification ability and data set difficulty are unlikely to be directly observable, taking inspiration from the kind of latent-variable models developed in psychometrics. I will also explore the value of causal explanations of observed performance of machine learning models and algorithms.

3.6 MagicWrangler Demo: Tool and Data

Jose Hernandez-Orallo (Technical University of Valencia, ES)

Joint work of Jose Hernandez-Orallo, Lidia Contreras-Ochando, Cèsar Ferri, Susumu Katayama, Fernando Martínez-Plumed, M. José Ramírez-Quintana
URL http://dmip.webs.upv.es/datawrangling/

MagicWrangler is a data wrangling tool that makes domain identification and extracts patterns using MagicHaskeller. We made a presentation of how the domain is identified and used to reduce the search space for MagicHaskeller. We presented a new data repository including 123 data wrangling datasets.

References

1 Contreras-Ochando, L.; Martínez-Plumed, F.; Ferri, C.; Hernández-Orallo, J.; and Ramírez-Quintana, M. J. "General-purpose inductive programming for data wrangling automation", AI4DataSci @ NIPS, 2016.

2 Contreras-Ochando, L.; Ferri, C.; Hernández-Orallo, J.; Martínez-Plumed, F.; Ramírez-Quintana, M. J.; and Katayama, S. "Domain specific induction for data wrangling automation (system demonstration)", AutoML @ ICML 2017.

3 Contreras-Ochando, L. "Domain specific induction for data wrangling automation", in Approaches and Applications of Inductive Programming (Dagstuhl Seminar 17382), Schmid, U. ; Muggleton, S. H. ; Singh, R. (eds.) Dagstuhl Reports, Volume 7, Issue 9. 2018, DOI: 10.4230/DagRep.7.9.86

4 Contreras-Ochando, L.; Ferri, C.; Hernández-Orallo, J.; Martínez-Plumed, F.; Ramírez-Quintana, M. J.; and Katayama, S. "General-purpose declarative inductive programming with domain specific background knowledge for data wrangling automation", https://arxiv.org/abs/1809.10054

3.7 Mapping the Skills in Data Science with Those in AI/ML

Jose Hernandez-Orallo (Technical University of Valencia, ES) and Lidia Contreras-Ochando

Joint work of Jose Hernandez-Orallo, Cèsar Ferri, Flach, P., Kull, M., Lachiche, N., Fernando Martínez-Plumed, M. José Ramírez-Quintana

Many skills, knowledge, abilities and competences have been identified as necessary for data scientists, including both technical and non-technical skills. Some standards, such as CRISP-DM (and extensions) recognise processes, but not competences or skills. In this talk, under the more general context of automation in the workplace, we explored how we can identify and map the skills in data science with the capabilities of AI and ML, in order to know when tasks can be effectively semi-automated in data science, according to the skills they need and the foreseeable progress in AI/ML.

We touched upon issues such as the value of semi-automation in front of whole automation, the problems of task 'atomising' instead of task automating, the issues of generality and autonomy in automation [3], the new needs for further supervision whenever automation takes place, and the overall view of automating data science as an intrinsic part of it, which actually makes data science evolve and require new tasks, skills and abilities.

We analysed what data science needs by first looking at the traditional data mining process (e.g., CRISP-DM), with a view of optimisation by considering a precise data mining goal, with then moving to more open data science trajectories [2], where data scientists require many non-technical skills and attitudes, and especially knowledge about the domain. We saw that more efforts for competence frameworks are needed, as what we have it is still too preliminary to characterise what data science requires in terms of skills.

On the other hand, we covered what AI/ML is providing, to see whether it is ready (now or in the near future) to cover partial automation in data science. We saw that apart from many other techniques in computer science, AI/ML provide solvers and learners, using Geffner's terminology [4]. If we restrict to ML only, there are some recent proposal of rubrics to see whether a task is automatable according to them, such as Brynjolfsson and Mitchell's rubric [1], which could be used for data science too. If we focus on the automatic handling of domain knowledge, there has been significant progress (e.g., NELL, Watson, etc.), but this is still poorly integrated with some other techniques.

Overall, we see that performing a well-informed mapping between what data science requires and what AI/ML can provide in the years to come is extremely challenging, perhaps more challenging than other areas where automation is playing an increasing role (e.g., transportation). This may be caused by the exploratory character of data science, which in the end is associated with the scientific methodology (data scientists are a kind of scientists). Nevertheless, we claimed that performing an analysis in terms of abilities or skills [5] will be more powerful and predictive in terms of degrees and opportunities for automation than just performing the analysis in terms of tasks.

References

1 Brynjolfsson, E. and Mitchell, T. "What can machine learning do? Workforce implications". Science, 358(6370):1530–1534, 2017.

2 Contreras, L., Ferri, C., Flach, P., Kull, M., Hernandez-Orallo, J. Lachiche, N., Martínez-Plumed, F. Ramírez-Quintana, M. J., "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories" in preparation.

3 Fernández-Macías, E., Gómez, E., Hernández-Orallo, J., Loe, B.S., Martens, B., Martínez-Plumed, F. and Tolan, S., 2018. A multidisciplinary task-based perspective for evaluating the impact of AI autonomy and generality on the future of work. arXiv preprint arXiv:1807.02416, AEGAP@IJCAI 2018.

4 Geffner, H. "Model-free, Model-based, and General Intelligence", htpps://arxiv.org/abs/1806.02308, 2018.

5 Hernández-Orallo, J. "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement", Artificial Intelligence Review 48 (3), 397-447, 2017.

3.8 A Taxomomy of Methods for Explainable Machine Learning Models

Tobias Jacobs (NEC Laboratories Europe – Heidelberg, DE)

In a short talk at Dagstuhl I have presented a taxonomy to categorize methods that contribute to explainable (including interpretable) machine learning models. A first fundamental dimension of the taxonomy is the purpose for which the method is applicable. Potential purposes range from technical and scientific benefits (e.g. model debugging, sanity checking, generating new insights) to societal requirements (e.g. validation of fairness, granting of of legal rights). The second dimension distinguishes between explainability as a model requirement or constraint during model construction on the one hand, and methods that help to explain existing complex models on the other. The latter kind of methods can be further categorized as methods which explain black-box models in terms of generic properties of the black box, or as methods to open the box and analyze what is happening inside. The final dimension distinguishes between methods to explain a model as a whole (also known as interpretability or global explainability) and methods to explain specific results of the model (also known as local explainability, or explainability in the narrow sense).

3.9 Counterfactual Prediction with Instrumental Variables and Deep Learning

Kevin Leyton-Brown (University of British Columbia – Vancouver, CA)

Joint work of Jason Hartford, Greg Lewis, Kevin Leyton-Brown, Matt Taddy
Main reference Jason S. Hartford, Greg Lewis, Kevin Leyton-Brown, Matt Taddy: "Deep IV: A Flexible Approach for Counterfactual Prediction", in Proc. of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, Proceedings of Machine Learning Research, Vol. 70, pp. 1414–1423, PMLR, 2017.
URL http://proceedings.mlr.press/v70/hartford17a.html

Counterfactual prediction requires understanding causal relationships between so-called treatment and outcome variables. This paper provides a recipe for augmenting deep learning methods to accurately characterize such relationships in the presence of instrument variables (IVs) – sources of treatment randomization that are conditionally independent from the outcomes. Our IV specification resolves into two prediction tasks that can be solved with deep neural nets: a first-stage network for treatment prediction and a second-stage network whose loss function involves integration over the conditional treatment distribution. This Deep

IV framework allows us to take advantage of off-the-shelf supervised learning techniques to estimate causal effects by adapting the loss function. Experiments show that it outperforms existing machine learning approaches.

3.10 Subjective Interestingness in Data Mining

Jefrey Lijffijt (Ghent University, BE)

Joint work of Jefrey Lijffijt, Tijl De Bie
URL http://users.ugent.be/ jlijffij/presentations/DAGSTUHL2018.pdf

I present a brief introduction to the topic of subjective interestingness, particularly an information-theoretic view that enables ranking of any type of patterns that we may want to extract from data. After this introduction, we review two ‘instances’ of this approach, for relational patterns (a generalisation of itemsets/tiles), as well as automatically finding informative views of data by seeing visualisations as patterns. Finally, I conclude with a question regarding to the topic of the workshop: are there fundamental or important differences between the topics of explainability and interpretability of machine learning models versus deriving insights from data.

Most of the talk covers material from the tutorial that we recently presented on this topic: http://www.interesting-patterns.net/forsied/tutorial/

References

1 Tias Guns, Achille Aknin, Jefrey Lijffijt, Tijl De Bie. “Direct mining of subjectively interesting relational patterns”. In Proceedings of the IEEE International Conference on Data Mining (ICDM), pp 913 – 918, 2016.

2 Bo Kang, Jefrey Lijffijt, Raúl Santos-Rodríguez, Tijl De Bie. “SICA: Subjectively Interesting Component Analysis”. Data Mining and Knowledge Discovery 32(4): 949-987, 2018.

3 Bo Kang, Kai Puolamäki, Jefrey Lijffijt, Tijl De Bie. “A tool for subjective and interactive visual data exploration”. In Proceedings of the European Conference of Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML-PKDD) – Part III, pp. 3 – 7, 2016.

4 Jefrey Lijffijt, Eirini Spyropoulou, Bo Kang, Tijl De Bie. “P-N-RMiner: A generic framework for mining interesting structured relational patterns”. International Journal of Data Science and Analytics, 1(1): 61-76, 2016.

5 Kai Puolamäki, Bo Kang, Jefrey Lijffijt, Tijl De Bie. “Interactive visual data exploration with subjective feedback”. In Proceedings of the European Conference of Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML-PKDD) – Part II, pp. 214 – 229, 2016.

3.11 AutoML Challenges 2015-2018: Review and Call for Action

Zhengying Liu (University of Paris Sud – Orsay, FR)

Joint work of Zhengying Liu, Isabelle Guyon, Hugo Jair Escalante, Michèle Sebag, Wei-Wei Tu, Andre Elisseeff, and numerous, other collaborators
Main reference Isabelle Guyon, Kristin Bennett, Gavin C. Cawley, Hugo Jair Escalante, Sergio Escalera, Tin Kam Ho, Núria Macià, Bisakha Ray, Mehreen Saeed, Alexander R. Statnikov, Evelyne Viegas: "Design of the 2015 ChaLearn AutoML challenge", in Proc. of the 2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12-17, 2015, pp. 1–8, IEEE, 2015.
URL https://doi.org/10.1109/IJCNN.2015.7280767

We introduce a series of data challenges in the research field of AutoML, organized by ChaLearn and many other organizations, with the support of numerous collaborators. These challenges are:

- AutoML challenge (2015-2016), collocated with NIPS, ICML, IJCNN;
- AutoML2 (2017-2018), collocated with PAKDD18;
- AutoML3: AutoML for Lifelong Machine Learning (on-going), collocated with NIPS18;
- AutoDL challenge (coming soon), more details to be announced.

References

1 Isabelle Guyon, Kristin Bennett, Gavin Cawley, Hugo Jair Escalante, Sergio Escalera, Tin Kam Ho, Núria Macia, Bisakha Ray, Mehreen Saeed, Alexander Statnikov, et al. Design of the 2015 ChaLearn AutoML challenge. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8. IEEE, 2015.

2 Isabelle Guyon, Lisheng Sun-Hosoya, et al. Analysis of the AutoML challenge series 2015-2018. In https://www.automl.org/book/

3.12 Bilevel Programming for Hyperparameter Optimization and Meta-Learning

Paolo Frasconi (University of Florence, IT)

Joint work of Luca Franceschi, Massimiliano Pontil, Paolo Frasconi, Riccardo Grazzi, Saverio Salzo
Main reference Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, Massimiliano Pontil: "Bilevel Programming for Hyperparameter Optimization and Meta-Learning". In Proc. of the 35th International Conference on Machine Learning ICML 2018, Stockholm, Sweden, July 10-15, 2018, pp. 1563–1572.
URL http://proceedings.mlr.press/v80/franceschi18a.html

We introduce a framework based on bilevel programming that unifies gradient-based hyperparameter optimization and meta-learning. We show that an approximate version of the bilevel problem can be solved by taking into explicit account the optimization dynamics for the inner objective. Depending on the specific setting, the outer variables take either the meaning of hyperparameters in a supervised learning problem or parameters of a meta-learner. We provide sufficient conditions under which solutions of the approximate problem converge to those of the exact problem. We instantiate our approach for meta-learning in the case of deep learning where representation layers are treated as hyperparameters shared across a set of training episodes. In experiments, we confirm our theoretical findings, present encouraging results for few-shot learning and contrast the bilevel approach against classical approaches for learning-to-learn.

3.13 Pyconstruct: A Library for Declarative, Constructive Machine Learning

Andrea Passerini (University of Trento, IT)

Joint work of Paolo Dragone, Stefano Teso, Andrea Passerini
Main reference Paolo Dragone, Stefano Teso, Andrea Passerini: "Pyconstruct: Constraint Programming Meets Structured Prediction", in Proc. of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden., pp. 5823–5825, ijcai.org, 2018.
URL https://doi.org/10.24963/ijcai.2018/850

Constructive learning is the task of learning to synthesize structured objects from data. Examples range from classical sequence labeling to layout synthesis and drug design. Learning in these scenarios involves repeatedly synthesizing candidates subject to feasibility constraints and adapting the model based on the observed loss. Many synthesis problems of interest are non-standard: they involve discrete and continuous variables as well as arbitrary constraints among them. In these cases, widespread formalisms (like linear programming) can not be applied, and the developer is left with writing her own ad-hoc solver. This can be very time consuming and error prone. I will describe Pyconstruct [1], a Python library tailored for solving real-world constructive problems with minimal effort. The library leverages max-margin approaches to decouple learning from synthesis and constraint programming as a generic framework for synthesis. Pyconstruct enables easy prototyping of working solutions, allowing developers to write complex synthesis problems in a declarative fashion in few lines of code. The library is available at: https://goo.gl/U1PaKF

References

1 Paolo Dragone, Stefano Teso and Andrea Passerini. Pyconstruct: Constraint Programming Meets Structured Prediction. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 5823–5825, 7 2018.

3.14 Pairwise Meta Rules, Full Model Selection, and Some Speculative Ideas

Bernhard Pfahringer (University of Waikato, NZ)

Joint work of Quan Sun, Mike Mayo, Bernhard Pfahringer
Main reference Sun, Q. (2014). Meta-Learning and the Full Model Selection Problem (Thesis, Doctor of Philosophy (PhD)). University of Waikato, Hamilton, New Zealand. Retrieved from https://hdl.handle.net/10289/8520
URL https://hdl.handle.net/10289/8520

In this talk I present pairwise meta rules for meta-learning, as well as a more scalable hierarchical version of them. They show good potential, especially when used with nearest neighbour, or random ranking forests. I also quickly cover an early attempt at an evolutionary system that jointly optimises combinations of up to four preprocessing methods and one learning algorithm. System Fantail combines ideas from genetic algorithms with particle swarm optimisation. Finally I speculate about approaches beyond black box optimisation.

3.15 Interactive / Visual Data Exploration Tutorial

Kai Puolamäki (University of Helsinki, FI) and Remco Chang (Tufts University – Medford, US)

We give an overview of interactive and visual data exploration. First, we discuss the scope of the talk, which include the analytic problems that benefit from both automation and human feedback, and in which the interaction happens at time scales of c. 1-10 seconds. We argue that practitioners would benefit from studying visualization techniques developed in vis community, such as graph techniques and techniques for multivariate data. Finally, we give an high level overview of dimensionality reduction and how to incorporate interaction there, plus a brief tutorial of how to get started on interactive and visual data exploration using R and Shiny library.

3.16 Tell Me Something I Don't Already Know: Tools for Human-guided Data Analysis

Kai Puolamäki (University of Helsinki, FI)

The outcome of the explorative data analysis (EDA) phase is vital for successful data analysis. EDA is more effective when the user interacts with the system used to carry out the exploration. A good EDA system has three requirements: (i) it must be able to model the information already known by the user and the information learned by the user, (ii) the user must be able to formulate the objectives, and (iii) the system must be able to show the user views that are maximally informative about desired features data that are not already know for the user. Furthermore, the system should be fast if used in interactive system. We present the Human Guided Data Exploration framework which satisfies these requirements and generalises previous research. This framework allows the user to incorporate existing knowledge into the exploration process, focus on exploring a subset of the data, and compare different complex hypotheses concerning relations in the data. The framework utilises a computationally efficient constrained randomisation scheme. To showcase the framework, we developed a free open-source tool, using which the empirical evaluation on real-world datasets was carried out. Our evaluation shows that the ability to focus on particular subsets and being able to compare hypotheses are important additions to the interactive iterative data mining process.

In this talk we present some tools for human-guided data analysis that utilise maximum entropy and/or constrained randomisation methods, such as sideR, available at http://www.iki.fi/kaip/sider.html, and tiler available at https://github.com/aheneliu/tiler

References

1 Puolamäki, Oikarinen, Atli, Henelius. Human-guided data exploration using randomization. arXiv:1805.07725 [stat.ML]

2 Henelius, Oikarinen, Puolamäki. Tiler: Software for Human-Guided Data Exploration. In Proc ECML-PKDD 2018, to appear. https://youtu.be/fqKLjMwJHnk

3 Puolamäki, Oikarinen, Kang, Lijffijt, De Bie. Interactive Visual Data Exploration with Subjective Feedback: An Information-Theoretic Approach. In Proc ICDE 2018, to appear. arXiv:1710.08167 [stat.ML]

3.17 Explaining Learned Models – Towards Relating Verbal Explanations to Visual Domains for Interactive Learning with Mutual Explanations

Ute Schmid (Universität Bamberg, DE)

Main reference Johannes Rabold, Michael Siebers, Ute Schmid: "Explaining Black-Box Classifiers with ILP – Empowering LIME with Aleph to Approximate Non-linear Decisions with Relational Rules", in Proc. of the Inductive Logic Programming – 28th International Conference, ILP 2018, Ferrara, Italy, September 2-4, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 11105, pp. 105–117, Springer, 2018.
URL https://doi.org/10.1007/978-3-319-99960-9_7

With the ever growing interest in machine learning in application domains such as autonomous driving, medical diagnosis, connected industry, or education, it has been recognized that machine learned models need to be transparent and comprehensible. For instance, a medical expert has to understand why a machine classifies some health state as critical given a computer tomographic image before s/he decides on the diagnosis. This is especially the case when the expert opinion deviates from the classification. In this case, the expert might (maybe wrongly) suspect a false alarm, mistrust the system, and consequently follow his/her own opinion. In the context of the collaborative project 'Transparent Medical Expert Companion' we are developing an approach to explanation generation for medical image data. Explanation generation is realized by a template based transformation of Prolog rules into natural language text. The Prolog rules has been learned with an inductive programming approach (Aleph, Metagol). Current topics of research in our project are: Combining black-box machine learning, especially Convolutionary Networks with ILP to address the trade-off between predicitive accuracy and interpretability; investigating incremental learning to allow the human experts to correct classifications which results in an adaptation of the classification rules; relating verbal explanations with the original image data, especially for explaining rules involving binary relations or negation.

3.18 Monte Carlo Tree Search for Algorithm Configuration: MOSAIC

Michele Sebag (CNRS, FR)

Joint work of Michele Sebag, Herilalaina Rakotoarison

The sensitivity of machine learning algorithms w.r.t. their hyper-parameters and the difficulty of finding the ML algorithm and its hyper-parameter setting best suited to the dataset at hand has led to the rapidly developing field of automated machine learning (AutoML), at the crossroad of meta-learning and structured optimization. Several international AutoML challenges have been organized since 2015, motivating the development of the Bayesian optimization-based approach Auto-Sklearn (Feurer et al. 15) and the randomized search approach Hyperband (Li et al. 16). In this paper, a new approach, called Monte Carlo Tree Search for Algorithm Configuration (MOSAIC), is presented, fully exploiting the tree structure of the algorithm portfolio-hyperparameter search space, with competitive results on the AutoML challenge 2015.

3.19 Statistical Thinking and Data Science: Observations

Padhraic Smyth (University of California – Irvine, US)

In the context of the seminar topic "automating data science" it is worth visiting potentially relevant ideas from the field of statistics, given the historical experience with the broad canvas of data analysis in that field. There has long been an acknowledgement in statistics of the importance of statistical thinking and statistical strategy in terms of providing a view of data analytic activities at a more general level than the details of specific methods and techniques. From a computer science perspective there is value in understanding how statisticians have addressed the question of how to think about data analysis in a systematic manner. This systematic perspective has been approached from the point of view of the practice of data analysis (see for example position papers by Tukey, Mallows, Cox, Huber, Chatfield, and many others), to the development of general theories and models for the process of data analysis (e.g., see Wilde 1994; Grolemund and Wickham, 2014), to education (e.g., Breiman, 1984 (US Berkeley Tech Report); Horton and colleagues, 2014 onwards), to the development of software systems to provide guidance to data analysts along the path of data analysis (e.g., see Oldford and Peters, 1986; Lubinsky and Pregibon, 1988). In this talk we briefly discussed these threads of work and their potential relevance to current endeavours in data science. In particular, one important message from this prior work is the notion that the human analyst is central to the process of data analysis and, as a consequence, tools that support the human analyst (i.e., semi-automation) are more likely to be successful than tools that seek to fully automate data analysis without a human in the loop.

3.20 Explainable Interactive Learning

Stefano Teso (KU Leuven, BE)

Joint work of Stefano Teso, Kristian Kersting
Main reference Stefano Teso, Kristian Kersting: ""Why Should I Trust Interactive Learners?" Explaining Interactive Queries of Classifiers to Users", CoRR, Vol. abs/1805.08578, 2018.
URL https://arxiv.org/abs/1805.08578

Although interactive learning puts the user into the loop, the learner remains mostly a black box for the user. Understanding the reasons behind queries and predictions is important when assessing how the learner works and, in turn, justifiably establishing or revoking trust. This talk covers some recent work on integrating active learning with explainable machine learning, where the queries to the user are augmented with predictions and explanations thereof, and the user provides both labels and explanation corrections for improved directability and control.

3.21 Automatic Configuration of Stream Clustering Algorithms?

Heike Trautmann (Universität Münster, DE)

Joint work of Heike Trautmann, Matthias Carnein

Analysing data streams has received considerable attention over the past decades due to the widespread usage of sensors, social media and other streaming data sources. A core research area in this field is stream clustering which aims to recognize patterns in an unordered, infinite and evolving stream of observations. Clustering can be a crucial support in decision making, since it aims for an optimized aggregated representation of a continuous data stream over time and allows to identify patterns in large and high-dimensional data. A multitude of algorithms and approaches has been developed that are able to find and maintain clusters over time in the challenging streaming scenario.

However, benchmarking stream clustering algorithms is a nontrivial task although first studies already exist. Besides the well-known problem of choosing an appropriate performance measure in unsupervised learning, the most crucial challenge when applying stream clustering algorithms is the correct choice of parameter settings. Stream clustering algorithms usually have a multitude of interdependent parameters, both for the micro-clustering step as well as for the macro-clustering phase and are highly sensitive to these settings. Automated algorithm configuration techniques would require an appropriate learning phase on the one hand and in our scenario moreover would have to be able to deal with drifts or structural changes of the stream. The talk aims at initiating a fruitful discussion on the topic paving the way to automated algorithm configuration and selection approaches.

References

1 Carnein, M., & Trautmann, H. (2018). Optimizing Data Stream Representation: An Extensive Survey on Stream Clustering Algorithms. Business and Information Systems Engineering (BISE), 2018. (Accepted)

2 Carnein, M., Assenmacher, D., & Trautmann, H. (2017). An Empirical Comparison of Stream Clustering Algorithms. In Proceedings of the ACM International Conference on Computing Frontiers (CF'17), Siena, Italy, 361–365.

3.22 Automatic Bayesian Density Analysis

Isabel Valera (MPI für Intelligente Systeme – Tübingen, DE)

Making sense of a dataset in an automatic and unsupervised fashion is a challenging problem in statistics and AI. Classical approaches for density estimation are usually not flexible enough to deal with the uncertainty inherent to real-world data: they are often restricted to fixed latent interaction models and homogeneous likelihoods; they are sensitive to missing, corrupt and anomalous data; moreover, their expressiveness generally comes at the price of intractable inference. As a result, supervision from statisticians is usually needed to find the right model for the data. However, as domain experts do not necessarily have to be experts in statistics, we propose Automatic Bayesian Density Analysis (ABDA) to make density estimation accessible at large. ABDA automates the selection of adequate likelihood models from arbitrarily rich dictionaries while modeling their interactions via a deep latent structure adaptively learned from data as a sum-product network. ABDA casts uncertainty estimation at these local and global levels into a joint Bayesian inference problem, providing robust and yet tractable inference. Extensive empirical evidence shows that ABDA is a suitable tool for automatic exploratory analysis of heterogeneous tabular data, allowing for missing value estimation, statistical data type and likelihood discovery, anomaly detection and dependency structure mining, on top of providing accurate density estimation.

3.23 Making Smart Data Analytics available for SMEs

Andreas Wierse (SICOS BW GmbH – Stuttgart, DE)

One of the main tasks of SICOS BW is to support small and medium sized enterprises (SMEs) in the uptake of data analytics technology (smart data with emphasis on the generation of value for the company). Since SMEs usually do not have a lot of expertise in data analytics (mostly none at all), we developed so called "potential analysis"-projects, where researchers of KIT (Karlsruhe Institute of Technology) work with the company's data for a few weeks in order to find interesting patterns but also to let them experience the process that is necessary to perform successful data analytics. If we find interesting patterns in the end the company is often interested to apply analytics in their every day business.

On one side this creates a need for training in order to provide the employees with the necessary knowledge; this is usually a fairly difficult and time consuming process. On the other side it would be very helpful for the SMEs, if the technology they want to use were highly automated. This could ease the burden for the SMEs significantly, since more employees would be able to use the technology and the need for training could be released. In general high usability is crucial for data analytics methods and tools to be successful in the SME context.

The companies that SICOS BW supports come from different sectors, including machinery, production, plant engineering and construction, manufacturing as well as trade an information technology. Application cases are for example predictive maintenance, sales forecast, early identification of production anomalies (chemical industry), energy consumption estimates or text classification.

References

1 Andreas Wierse, Till Riedel. *Praxishandbuch Smart Data Analytics.* De Gruyter Oldenburg, Berlin, Germany, 2017.

3.24 Tutorial on Data Preparation and Cleaning

Christopher Williams (University of Edinburgh, GB)

Joint work of Christopher Williams, Taha Ceritli, James Geddes, Zoubin Ghahramani, Ernesto Jimenez-Ruiz, Ian Horrocks, Alfredo Nazabal, Tomas Petricek, Charles Sutton, and Gerrit Van Den Burg of the Artificial Intelligence for Data Analytics (AIDA) project at the Alan Turing
URL https://www.turing.ac.uk/research/research-projects/artificial-intelligence-data-analytics

A common view is that up to 80% of work on a data mining project is involved in data understanding, cleaning and preparation, yet machine learning has not focused very much on these topics. I will describe the issues around data parsing, obtaining (or inferring) a data dictionary, data integration, entity resolution, addressing format variability, structural variability, identifying and repairing missing data, and anomaly detection and repair.

4 Working groups

The following sections contain summaries of breakout discussion session that were organized during the seminar. The authors for each section led the discussion of one of these breakout sessions. They made an attempt to summarize these discussions in an accurate, comprehensive, and intelligible manner, fairly reflecting all points of view. Although it is not practical to attribute all input to individual participants, note that these summaries are thus the result of input by several, often many of the workshop participants. As such, the summaries also do not always reflect either the author's or the full breakout session group's opinion.

The workshop and the discussion organizers sincerely appreciate the input from all the participants in the breakout sessions and would like to thank everyone for their contributions.

4.1 Breakout session on "Beyond the Black Box – Explainability of Machine Learning Models"

Ray G. Butler (Butler Scientifics – Barcelona, ES)

The breakout session "Beyond the Black Box – Explainability of Machine Learning Models" discussed issues related to being able to explain and interpret the models that machine learning produces. We made a distinction between explainability – the ability of a model to justify decisions/predictions in individual cases – and interpretability – understanding a model as a whole and in general, perhaps through a series of decisions/predictions (a "conversation" with the model). In particular, explainability is relatively easy to achieve and already present for most models (for example each prediction can be traced through a series of neurons that represent certain features in a deep neural network), while interpretability is poorly understood even for simple models. An interpretable model would allow to judge any and all biases that were present in the training data and the model to assess whether the model will work as expected in every case.

An explanation or interpretation of a model is subjective in that one human may be satisfied with/convinced by a particular explanation that another human finds unsatisfactory.

There is literature in the social sciences on what explanations are and how humans perceive them that should be helpful in this context. A caveat to watch out for when providing explanations is that humans are very good at matching patterns and convincing themselves that something makes sense, and an incorrect or misleading explanation that is not supported by the data may sound convincing.

Finally, we discussed the issue of performance in this context. While in many cases only complex and hard-to-interpret models deliver good performance, the use of AutoML methods can help in that it allows to train even relatively simple models with good performance. Explainability can also help to increase performance, as it allows human experts to inspect the model, identify areas where performance is bad and the reasons for that, and remedy those issues.

4.2 Breakout Session on "Automation of data exploration tasks"

Tijl De Bie (Ghent University, BE)

Many important data science tasks are ill-specified: there is some amount of data (increasingly an abundance of it, and often heterogeneous and with complex structure), and the data analyst wishes to gain new insights or make discoveries driven by this data. The main related challenges stem from the fact that such tasks are purposefully ill-defined and open-ended – i.e. there is no clear objective function like in more traditional machine learning tasks where the goal is typically to build a predictive model for a specific target variable.

The key challenge in automating data exploration is thus how to specify and pursue an inherently ill-defined objective, and how to measure progress towards such an objective.

It was mentioned that data exploration is often an intermediate step in a data science pipeline, the end goal of which is well-defined but too complex to tackle without a good understanding of the data. An example was given of investment companies gathering large amounts of data, which is first analyzed in an exploratory manner by data profilers, before it is used for predicting stock price fluctuations – the ultimate goal which is predictive. A more elementary example is the well-known use of of feature selection or dimensionality reduction as a regularization strategy for a subsequent machine learning model.

However, several participants raised examples where data exploration is arguably the end-point. A first example given is data-driven research such as (increasingly) biological research. The end goal of biological research is arguably to understand how life works. This is a scientific discovery problem for which researchers increasingly rely on high-throughput data – i.e. it is increasingly addressed using data science techniques. Of course, it is inevitable to break this task down into smaller subproblems (e.g. "does this particular subsequence in the DNA code for a gene or not?"), but these are merely steps towards the larger goal of using data in order to gain an understanding of life. Other examples that were raised pertain to astronomy research (e.g. the Sloan Digital Sky Survey initiative), as well as to industry (e.g. the placement of a multitude of sensors throughout an industrial plant, which may be used to find anomalies in the process and understand the process better).

Thus, a distinction can be made between tasks where a clear objective exists although it is too complex in practice to pursue without an initial exploratory analysis of the data, and tasks where such an objective function does not exist at all and the goal is simply exploring the data in order to make data-driven discoveries or gain new insights. These different settings may have different needs in practice, and thus have been studied in different communities (e.g. the databases, the statistics and exploratory data analysis, and the KDD communities) using different kinds of techniques. Arguably, the roots are to be found in statistics, and in the seminal work of John Tukey on Exploratory Data Analysis (EDA), who suggested EDA as a way of hypothesis generation, to be subsequently subjected to confirmatory analyses.

A point of discussion was whether the automation of data exploration can be approached in a similar manner as AutoML techniques for supervised learning. To an extent this seems to be the case:

- Where the exploratory analysis is merely an intermediate step, automating this step can at least in principle be driven by the quantifiable objective of the entire data science process (which could be a predictive modeling task). Yet, in practice this is typically infeasible.
- There are also similar issues of model and/or algorithm selection to be done in prototypical data exploration tasks such as outlier detection, clustering, community detection in networks, dimensionality reduction, and density estimation. In those tasks, some notion of model fit (e.g. the likelihood of a model, a clustering cost function, etc.) is an obvious criterion. However, such a notion may not always align with the needs of the human data analyst (e.g. it may explain irrelevant aspects of the data very well, at the expense of the more relevant aspects in the data). Thus, user interaction seems fundamentally inevitable to this process, whereas in AutoML user interaction is arguably needed only due to imperfections in AutoML that at least in principle could be remedied with more research progress.

Some other orthogonal issues were raised in the discussion, and briefly summarized here.

The first remaining issue is the need for methods that can extract insights from and make discoveries in structured data. Many techniques still expect the data to be formatted in a basic data table, often causing significant loss of information.

The second remaining issue is the need to take into account background knowledge, in two ways. The first way is in ensuring that the findings in the data add new value to that prior background knowledge, i.e. they should not be implied by it as the analyst would not gain new insights. An approach like this would ensure the findings are subjectively interesting to the data analyst. The second way is almost the opposite: ensuring that the findings in part corroborate the prior knowledge of the data analyst, in order to create trust in the analysis. Hence, there is a trade-off between trust and information to be made.

The third remaining issue concerns visualization, often a crucial part of data exploration. Quantifying the quality of a visualization is an open challenge. A particular aspect not often studied is the veracity of visualizations, or lack thereof: the risk that a visualization can make the data analyst see patterns that are not there (e.g. t-SNE is known to exhibit spurious cluster structure if the parameters are not tuned right). To conclude, the exploration of data was acknowledged as an important data science component, sometimes as a stepping stone towards a clearly defined goal, and sometimes as an ill-specified but important goal on itself. Such tasks are challenging in practice, and considerable amounts of research are needed to automate it further – but inevitably with the user tightly in the loop.

4.3 Breakout Session on “Data Wrangling”

Jose Hernandez-Orallo (Technical University of Valencia, ES)

We had a breakout session on “Data Wrangling” on Tuesday afternoon. We understood the term as covering all elements in data preparation, such as data integration, cleansing and transformation, e.g., following Chris William’s taxonomy, presented in his tutorial. We wanted to make this clear as some definitions of data wrangling are too narrow (e.g., https://en.wikipedia.org/wiki/Data_wrangling).

During the discussion we covered some of the data wrangling topics that were considered most relevant during the previous days. In particular we covered:

- How to instruct the system on what we want to do with data? (‘domain specific wrangling’): Given a dataset and a task (for which we are given, in the best cases, a performance metric) can we clean (apply operators to) the dataset to maximise the performance metric? We understand cleaning here as going beyond feature extraction. Or, given a dataset and a target algorithm, can we put the data in the right format for the particular target algorithm? An issue is whether transformation are reversible or traceable (need to be able to come back to the original representation). One question that was raised is whether data wrangling can be done just to increase “data quality” without having a particular task in mind. For instance, domain inference can be done and is useful even if we do not know the task yet. When talking about “domain specific wrangling”, we saw the important distinction between the format (more task-independent) and the semantic part (more task-dependent).
- Automatic data integration: this item, which is related to the previous one, considers the integration of different sources of data, in different formats. Can we have a sequence of data wrangling operations that lead to a desirable format? Again, we are here very interested in ensuring provenance (an issue not only but mostly for integration).
- How can we make tools more general where users can add their knowledge instead of being preprogrammed? We mentioned some possibilities, such as the use of inductive logic programming, for which this is natural. As an example, Pyconstruct is a library for declarative, constructive machine learning (Andrea Passerini) where background knowledge is expressed as constraints. One question that arose is whether we need domain-specific ontologies or domain adaptation (use knowledge/models from different but related domains). Finally, the interfaces for interaction are most relevant here, and learning can be transformed into teaching: using “teaching by example” followed by program synthesis (e.g., FlashFill).
- Quantifiable measurable progress and metrics for data wrangling. We all agreed that we need benchmarks, for exploratory settings (with no predetermined task) and for settings where there is precise goal. We need metrics that estimate how close the data wrangling process gets us to some format that could be used in a data analysis tool (weka, sk-learn, Knime, ...). For the evaluation in supervised settings, it is important not to apply the semantic transformations to the test data (e.g., the test data should not use instance selection motivated by missing values).
- Non-tabular data (text, sensor, video, etc.). Most of the data wrangling literature thinks in terms of converting data into one single integrated minable view, in a tabular way, with columns (features) and rows (examples). But many problems do not fit this setting. Some may use embeddings to convert between representations, especially in NLP, but in other

cases the representations are created (possibly automatically by learning representations techniques), and both the format and the domain may change with time and must require adaptation.

- Feedback from the users: finally, and related to many of the items above, it is very important for semi-automated approaches or for improving and validating the fully-automated ones to have feedback from the users. For instance, Flashfill is an example of simple feedback, but there are other ways of feedback (reinforcement learning, preferences, such as collaborative filtering, etc.). Another point that was made is that intervening early in the process might be easier and more effective.

As possible actions from this workshop we suggested the collection of related papers, the analysis of what ETL / DS tools provide nowadays, having a look at sites such as “frictionless data” https://frictionlessdata.io/ including csv files and code for wrangling, the derivation of new metrics and collection of datasets. As a more long-term action we talked about a possible challenge/competition.

4.4 Breakout session on “Data-driven + / vs knowledge-based techniques (learning + / vs automated reasoning)”

Andrea Passerini (University of Trento, IT)

The discussion started by pointing out that automated reasoning is being largely overlooked nowadays because of the hype on machine learning and especially deep learning. However, there is a strong potential in the combination of machine learning techniques and automated reasoning / knowledge-based approaches, which is definitely worth exploring.

Three main ways of combining these fields were identified:

- using machine learning to improve reasoning systems
- using reasoning techniques to improve machine learning systems, e.g. by adding a reasoning layer on top of machine learning models
- learning to reason

One non-trivial aspect which had to be addressed was what is meant by knowledge. Generally speaking, any model made out of data could be considered knowledge. The consensus was that we talk about knowledge meaning explicit knowledge, see also the distinction between explicit and implicit knowledge in psychology.

A second relevant aspect was that the distinction between learning and reasoning does not imply a distinction between data-driven and knowledge-driven approaches, as learning itself can be both data-driven and knowledge-driven, and the same holds for reasoning. The distinction between data-driven and knowledge-driven approaches should thus be complemented with the distinction between induction and deduction.

In terms of usefulness of the combination, these are the main opportunities we identified:

- More accurate models
- More efficient models (one-shot learning)
- Reliable models
- Interpretable models
- Augmenting humans (fix reasoning biases, improve reasoning capabilities)

When talking about more developing accurate models, the field of statistical relational learning is an obvious candidate. However, many questions remain unanswered:

- Do we need statistical relational learning, when plenty of data is available?
- The computational cost of these systems is substantially higher that the one of deep learning systems (also thanks to advances in hardware), can this gap be filled in some way?

When talking about interpretable models and augmenting humans, the need for understanding the principles and limitations of human reasoning clearly emerged as a necessary and poorly explored aspect (see e.g. Thinking, fast and slow by Daniel Kahneman). From this perspective, the fact that existing approaches have mostly focused on propositional and first order logic seems suboptimal or at least incomplete. Other modalities should be better explored, like:

- argumentation
- confirmation theory
- causality
- counterfactual reasoning
- conversational reasoning

When talking about reliability of learning systems, while explainability clearly helps in building trust, it is not always required:

- Technologies get accepted and used, even if we do not fully understand them.
- Unsat proofs from SAT solvers can be huge (but they can be verified by simple proof checkers)
- We trust black boxes all the times, they are called humans!
- There are two different kinds of trust – the one that leads us to trust in people and the one that lets us to trust engineered systems (planes, etc.)

4.5 Breakout Session on "Human-in-the-Loop Automatic Data Science, and How to Avoid Ignorant Use"

Joaquin Vanschoren (TU Eindhoven, NL) and Holger H. Hoos (Leiden University, NL)

An important question arising in the context of automating data science concerns the degree of automation that is feasible or desirable. This breakout session started from a broad agreement that complete automation is currently infeasible as well as conceptually problematic. Two groups explored and discussed the concept of human-in-the-loop AutoDS, i.e., automated data science that supports human users with a certain level of expertise rather than aiming to replace them, and how to avoid ignorant use of human-in-the-loop AutoDS systems.

4.5.1 Human-in-the-loop AutoDS

Often, a user does not know precisely what she wants. To effectively aid such users, it seems inevitable to include user modelling in the system: the AutoDS system should learn to understand the intentions and limitations of specific users. This leads to 'personalized data science': AutoDS systems should figure out what matters to their users, and how to best support them in achieving their objectives. Realising such capabilities also requires expertise in human-computer interaction. The objectives and goals of human users of an AutoDS system (or any data science tool) will be substantially affected by their level of knowledge. AutoDS systems should assess this, for example, by running experiments designed to assess a user's expertise and tendency to blindly trust the system. Especially for inexperienced users, advice such as 'people who used this model/technique also found that model/technique useful' may be helpful.

The existing research area of preference learning is highly relevant in this context. One useful approach would be for the AutoDS system to start processing given data and ask the user for input at carefully chosen points in time. Interesting research questions arising in this context include the following: Where in a data science pipeline can the user meaningfully / effectively add knowledge, and where is it most needed? Which kind of user feedback should be elicited if there is a limit on the number of interactions or the amount of overall time allotted to them? Specific questions to elicit user preferences are "Can you understand this model?", or "I have removed these outliers, was that OK?" There is a limit on the amount of feedback a user is willing to give, so the AutoDS system should be conscientious in requesting it. Considering that humans often work very effectively with visualisations, these should likely play a key role in such interactions.

Another mode of interaction that may be interesting to consider is that found in so-called centaur teams in chess playing, where a human player suggests a move to chess program (before actually making the move), and the program analyses and demonstrates what will likely happen. This type of interaction has been shown to lead to powerful interactions and strength of play beyond that of the best chess programs.

Overall, it seems that finding compelling answers to the questions arising in the context of human-in-the-loop systems requires a significant shift in research focus for many AutoDS researchers; for example, there is currently little interest in user studies. An open question is how to incentivise researchers to work on these questions.

4.5.2 How to avoid ignorant use of AutoML/AutoDS?

Ignorant use (in the sense of use without necessary understanding of key characteristics and limitations) of machine learning and data science methods can lead to poor performance, misleading results and ultimately, incorrect or harmful decisions. Considering the degree to which machine learning and data science techniques are starting to be used within organisations without access to the prerequisite expertise, serious problems are bound to arise, especially as increasing automation promises to make data science techniques more accessible to non-experts. A crucial question therefore is how to prevent or alleviate the problems arising from ignorant use as much as possible.

First, we need better education and training. Many online courses now promise to make anyone taking them into a data scientist with minimal effort and time investment. Worse, there is a misperception that AutoDS will soon completely eliminate the need for data scientists. It is important for experts in AutoDS to actively warn against this view. One idea is to generate counterexamples where AutoDS does not work yet, and share war stories to make people more aware of current pitfalls. It may also be useful to start a blog on how not to automate data science.

Second, we need to provide guardrails against ignorant use of AutoDS. For instance, users could be prevented from or warned against performing multiple comparisons without proper statistical correction. Yet, distributed reuse is difficult to control. Other preventive measures could include that learned models should refuse to make predictions if the inputs are too different from the training data. However, if the model is poor, this may be difficult to recognise. Overall, the idea of guardrails against ignorant use raises interesting research questions: What are conditions that should be met in order to safely use a method? Can we learn those? How can we best support users in meeting these conditions?

Generally, the automation of data science is likely subject to a generalised form of Wiener's laws of aviation and human error, specifically: Digital devices tune out small errors while creating opportunities for large errors.

With this in mind, and related to the second point above, it seems useful to distinguish two types of automation. The first is to eliminate drudgery, as in a washing machine that (for most purposes) eliminates the need for manual washing. The second is to provide high-level oversight to help detect, avoid or compensate for human error, as in an fly-by-wire system that would not accept control inputs from a pilot that would stall the aircraft. It appears that good progress is being made in the area of drudgery automation, i.e., in automating tasks that data scientists have to do routinely, such as hyperparameter optimisation. Less attention has been paid so far to the second type of automation, which may hold the key to avoiding many of the pitfalls of ignorant use of AutoDS systems.

Overall, it seems useful to bracket human expertise by the two types of automation, which we may call a "human in the centre" approach. To avoid complacency, it may also be useful to occasionally let humans perform tasks that can be automated (this has, in fact, be proposed in aviation to counteract the detrimental effects of autopilot use on human pilots' skills). Finally, it may be interesting to investigate to which degree an AutoDS system could recognise when it can automate tasks for a specific user, and how to do so safely, without eliminating or reducing important elements of human judgement.

It seems clear that, taking a the "human in the centre" approach, increasing automation from both sides will gradually reduce the role of the human user or operator. Whether this will ultimately make it possible to completely eliminate the need for a human expert is an open question; even if it were possible, it is unclear whether this would a desirable goal to achieve. As long as human experts play an important role in the data science process, it seems crucial to ensure that they have an appropriate level of knowledge and preparation that allows them to safely use the system – perhaps in the form of a meaningfully defined 'data science license'.

Participants

- Leman Akoglu
Carnegie Mellon University – Pittsburgh, US
- Mitra Baratchi
Leiden University, NL
- Michael R. Berthold
Universität Konstanz, DE
- Hendrik Blockeel
KU Leuven, BE
- Pavel Brazdil
University of Porto, PT
- Ray G. Butler
Butler Scientifics – Barcelona, ES
- Remco Chang
Tufts University – Medford, US
- Felipe Leno da Silva
University of São Paulo, BR
- Tijl De Bie
Ghent University, BE
- Luc De Raedt
KU Leuven, BE
- Peter Flach
University of Bristol, GB
- Paolo Frasconi
University of Florence, IT
- Elisa Fromont
University of Rennes, FR
- Jose Hernandez-Orallo
Technical University of Valencia, ES
- Holger H. Hoos
Leiden University, NL
- Frank Hutter
Universität Freiburg, DE
- Tobias Jacobs
NEC Laboratories Europe – Heidelberg, DE
- Lars Kotthoff
University of Wyoming – Laramie, US
- Nada Lavrac
Jozef Stefan Institute – Ljubljana, SI
- Kevin Leyton-Brown
University of British Columbia – Vancouver, CA
- Jefrey Lijffijt
Ghent University, BE
- Zhengying Liu
University of Paris Sud – Orsay, FR
- Siegfried Nijssen
UC Louvain, BE
- Andrea Passerini
University of Trento, IT
- María Pérez-Ortiz
University of Cambridge, GB
- Bernhard Pfahringer
University of Waikato, NZ
- Kai Puolamäki
University of Helsinki, FI
- Matteo Riondato
Two Sigma Investments LP – New York, US
- Ute Schmid
Universität Bamberg, DE
- Marc Schoenauer
INRIA Saclay, FR
- Michele Sebag
CNRS, FR
- Padhraic Smyth
University of California – Irvine, US
- Alexandre Termier
University Rennes, FR
- Stefano Teso
KU Leuven, BE
- Heike Trautmann
Universität Münster, DE
- Isabel Valera
MPI für Intelligente Systeme – Tübingen, DE
- Matthijs van Leeuwen
Leiden University, NL
- Joaquin Vanschoren
TU Eindhoven, NL
- Jilles Vreeken
Universität des Saarlandes, DE
- Andreas Wierse
SICOS BW GmbH – Stuttgart, DE
- Christopher Williams
University of Edinburgh, GB