EMBnet.journal
Bioinformatics in Action

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

Technical Notes

The CHARME "Advanced Big Data Training School for Life Sciences": an example of good practices for training on current bioinformatics challenges

Yen Hoang[1], Juliane Pfeil[2], Maja Zagorščak[3]✉, Axel Y. A. Thieffry[4], Eftim Zdravevski[5], Živa Ramšak[3], Petre Lameski[5], Sabrina K. Schulze[6], Eleni Papakonstantinou[7], Louis Papageorgiou[7,8], Tarry Singh[9], Ariel Duarte-López[10], Marta Pérez-Casany[11]

[1] German Rheumatism Research Center Berlin, A Leibniz Institute, Germany

[2] Division Molecular Biotechnology and Functional Genomics, Technical University of Applied Sciences, Wildau, Germany

[3] Department of Biotechnology and Systems Biology, National Institute of Biology (NIB), Ljubljana, Slovenia

[4] Department of Biology, Biotech Research & Innovation Centre, University of Copenhagen, Denmark

[5] Faculty of Computer Science and Engineering, Sts. Cyril and Methodius University in Skopje, Skopje, Macedonia

[6] Institute of Biochemistry and Biology, Cell2Fab, University of Potsdam, Potsdam, Germany

[7] Genetics and Computational Biology Group, Laboratory of Genetics, Department of Biotechnology, Agricultural University of Athens, Greece

[8] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece

[9] CEO and AI Neuroscience Researcher of deepkapha.ai Labs, Deepkapha.ai, Amsterdam, Netherlands

[10] Department of Computer Architecture, Technical University of Catalonia, Barcelona, Spain

[11] Department of Statistics and OR, Technical University of Catalonia, Barcelona, Spain

Competing interests: YH none; JP none; MZ none; AYAT none; EZ none; ŽR none; PL none; SKS none; EP none; LP none; TS none; ADL none; MPC none

Abstract

The CHARME "Advanced Big Data Training School for Life Sciences" took place during 3-7 September 2018, at the Campus Nord of the Technical University of Catalonia (UPC) in Barcelona (ES). The school was organised by the Data Management Group (DAMA) of the UPC in collaboration with EMBnet as a follow-up of the first CHARME-EMBnet "Big Data Training School for Life Sciences", held in Uppsala, Sweden, in September 2017. The learning objectives of the school were defined and agreed during the CHARME "Think Tank Hackathon" that was held in Ljubljana, Slovenia, in February 2018.
This article explains in detail the step forward the organisation of the training school, the covered contents and the interaction/relationships that thanks to this school have been established between the trainees, the trainers and the organisers.

Introduction

The "Advanced Big Data Training School for Life Sciences"[1] (ATS-LS) was a joint activity of the EU COST Action CHARME[2] (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research - CA15110) and EMBnet[3] (The Global Bioinformatics Network), the former providing financial and the latter organisational support.

After the first "Big Data Training School for Life Sciences", held on September 2017 in Uppsala, Sweden (Pfeil *et al.*, 2018), some participants showed interest in organising a more advanced version of the event. This idea, supported by the COST Action CHARME's management committee members, who collaborated to the organisation of the first school edition, was finalised at the CHARME "Think Tank Hackathon" that took place in Ljubljana, Slovenia (Schulze *et al.*, 2018) in February 2018.

Out of the 48 received applications, 23 were selected (including five local participants) to attend the school and only 16 selected by the scientific committee[4] to be awarded a stipend (CHARME grant). This selection

[1] http://dama-advancedbigdataschool.ac.upc.edu/

[2] http://www.cost-charme.eu/

[3] http://www.embnet.org/

[4] http://dama-advancedbigdataschool.ac.upc.edu/apply

Article history
Received: 02 October 2018
Accepted: 07 January 2019
Published: 30 January 2019

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

was based on specific criteria that were: i) the prior knowledge of programming and statistical methods in life sciences; ii) the participation of candidates at the first Big Data training school in Sweden; iii) the research background, and iv) candidate's motivation. Additional criteria were ethnicity, gender and national affiliation that were considered for keeping an overall balanced diversity.

The training school officially commenced on the 3rd of September, with opening words and a welcoming introduction from the vice-Chair of the CHARME action, Prof. Erik Bongcam-Rudloff, and the local organisers, Prof. Marta Pérez-Casany and Ariel Duarte-López. The programme encompassed three main frameworks composed of interactive lectures accompanied by hands-on sessions and roundtables in the categories of:

1. Feature Selection (FS)
2. Machine Learning (ML) using the Microsoft Azure platform and Virtual Machines (VMs)
3. Deep Learning (DL) and Artificial Intelligence (AI) applied to life science and healthcare data, leveraging the power of graphics processing units (GPUs)

The rest of this article is organised in six sections. The first three sections describe the organisation and scope of special lectures whereas the other ones present other essential aspects of the school, such as the "Trainers' Perspectives", a "Miscellaneous" section, that includes different information and a "Summary".

The section "Availability of data and materials" provides links to all materials, including presentations, code repositories, and datasets.

Topic I: Feature Selection

Andrzej Janusz, Assistant Professor, Institute of Informatics, University of Warsaw, Poland, opened the ATS with a one-day session, exhaustively introducing the topic of Feature Selection (FS) that covered:

- the terminology of feature and its synonyms;
- how FS reduces complexity, improves interpretability either for explicative as well as predictive models and generally contributes to having more robust models;
- basic FS tools that are used in linear models as the t-statistic;
- common misconceptions regarding FS;
- importance of understanding problem space, relevance measure, redundancy, and optimality;
- underlying assumptions, checklist and decision making, as well as the concepts of scalability, stability, generalisation, and overfitting;
- curse of dimensionality; the difference between dimensionality reduction methods and FS;
- categorisations of FS techniques from a perspective of data type, learning task and selection strategy; filter methods (univariate and multivariate), wrapper methods, embedded methods;
- decision redacts, bottom-up vs top-down approaches;
- optimisation algorithms, iterative methods, metaheuristics;
- discussion about why feature selection is still relevant in the dawn of DL.

At the end of the forenoon, students were presented with detailed walkthrough methods, some FS "traps" and "recipes". The afternoon topic was reserved for hands-on sessions followed by questions and answers, held in a round-table manner. Live coding was conducted in R (R Core Team, 2017) version 3.4.3 or higher using IRkernel and Jupyter Notebook (or Markdown, arbitrarily) with R packages of interest: ggplot2_3.0.0, FSelector_0.31, caTools_1.17.1.1, dprep_3.0, ROCR_1.0-7, data.table_1.11.4, digest_0.6.17, uuid_0.1-2, devtools_1.13.6, repr_0.15.0. Some of the packages above were archived from the CRAN repository due to uncorrected problems and required manual installation, which turned out to be a cumbersome process for an unpractised R user.

The first dataset of interest was a benchmark dataset related to breast cancer diagnostics. Acknowledging the importance of data visualisation before analysis, dimensionality reduction using principal component analysis was performed. The dataset was pre-processed, and usefulness of individual features was checked using various filtering approaches, *e.g.* t-test, Wilcoxon-test, and chi^2-test. Amongst others, the predictive power of individual attribute was assessed, by computing the area under the ROC curve (AUC) scores. To explore multi-class problems, multivariate feature ranking algorithms were applied. Those were implemented in several R libraries and fine-tuned according to exercise requirements. The decision on which attributes to choose was made upon multiple strategies (Riza *et al.*, 2017). One of the strategy in the form of R package, RmRMR, is publicly available from Andrzej Janusz's Github repository (https://github.com/janusza/). Furthermore, we explored various wrapper and embedded methods (*e.g.*, decision trees and neural networks). The wrappers search heuristics covered: sequential forward search, recursive feature elimination, hill-climbing, simulated annealing, and genetic algorithms.

To raise awareness of false consequences, random FS was applied to synthetic data and resulted in a good set of predictors, that however lead to misinterpretation of the model's generalisation. Students were motivated to repeat the experiment using a proper methodology, make predictions and compare the outcomes. Likewise, using synthetic datasets with different noise manipulations, the task was to find a suitable FS method with high performance.

The last dataset of interest was a real-world example: an acute lymphoblastic leukaemia microarray dataset. The purpose of the exercise was to independently classify cases of leukemia into subtypes using some of the discussed techniques. At the end of the hands-on session, appropriate solutions, broken down into steps/sub goals, were shown and discussed.

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

Topic II: Microsoft Azure data science and Machine Learning services of interest

This two-day topic began with a lecture by Dimitrios Vlachakis, Assistant Professor, Genetics Laboratory, Department of Biotechnology, Agricultural University of Athens, Greece, accompanied by two teaching assistants of his research group, Eleni Papakonstantinou, and Louis Papageorgiou. In the preface, Dimitrios gave a brief introduction to the Big Data topics and the way they are encountering in the fields of genomics, molecular dynamics, development of anticancer and antiviral agents, immunoinformatics and neurodegenerative diseases investigations. He justified the necessity of cloud computing for Big Data analysis and the various services offered by Microsoft Azure (*e.g.*, virtual machines), cloud services, websites, and mobile services. In preparation for this topic, all students were asked to sign in to Microsoft Azure Cloud. The local Organising Committee considered three ways to get free Azure credits. First-time users received 170 EUR credits and free 12-month use[5] to explore the different services on the Azure Cloud. Also, Alberto Marcos González, Higher Education Account Executive for Microsoft Spain, provided 25 vouchers with 100 USD (85 EUR) credit. As a third safety measure, Eftim Zdravevski, one of the student participants, kindly offered the students to use some credits from his research grant.

Under the technical supervision of Eleni Papakonstantinou, the students went through a practical introduction to Windows Azure Data Services, followed by a hands-on session of SQL database creation and data upload procedures. After the topic of handling big data with Azure, the idea was to work with Azure Machine Learning Studio (AMLS)[6]. However, the Microsoft servers were on a temporary shutdown due to external environmental influences. Meanwhile, Raik Otto, technical assistant for Topic III, attempted to lead the students through preparation of Azure GPU NC-series VM instances, mandatory for future tasks. This was also affected by the servers' outage, so Dimitrios' fast improvisation skills came to the fore, his pharmacogenomics presentation allowing an uncompromised continuation of the learning experience.

Later that night, it was publicly announced that Azure was accessible again. According to the official webpage[7] section "9/4/2018 South Central US", there was an internal error of all Azure resources, which made it impossible to work with Azure. It was caused by a high energy storm that hit the Azure data centre before it could safely shut down. A significant number of storage servers were damaged, as well as a small number of network devices and power units.

[5]https://azure.microsoft.com/en-us/free/

[6]https://studio.azureml.net/

[7]https://azure.microsoft.com/en-us/status/history

At the beginning of the next day, Raik Otto once more led the way in preparations for Topic III. This time the alternative Azure account creation went through successfully. Thereupon the student pairs applied for an Azure pass[8], receiving an additional 85,00 EUR of free credits and the ability to request for an NC6 series instance, i.e., a GPU optimised virtual machine suited for deep learning problems. Some participants with capabilities to run their resources also prepared those by installing Linux Ubuntu (16.04 or higher); the NVIDIA graphic card driver (user specific), additional Nvidia libraries (cuDNN, NCCL), conda, python 3.6, openjdk, bazel, TensorFlow (from sources), Keras and Jupyter Notebook. To avoid potential problems and challenges, Eftim Zdravevski and Petre Lameski, Assistant professors at the Faculty of Computer Science and Engineering, University Sts. Cyril and Methodius in Skopje, Macedonia, prepared computational resources of their faculty, to be shared with participants in case of a need.

Successful infrastructure setup was smoothly boosted by a short talk about MS Azure by Alberto Marcos González, Higher Education Account Executive, Microsoft, and followed by the hands-on session under Eleni's supervision. She introduced AMLS and few "how-to's", graphic user interface that toggles focus from coding to interactive data analysis workflows and visualisation as the machine learning procedures adapted to newbies and experts. Machine learning experiment was built from scratch using "Wisconsin Breast Cancer Diagnostics" dataset. Data were accordingly processed, visualised and summarised. Feature selection procedure was applied, and outputs from different ML algorithms were compared. Models were scored and evaluated, which was followed by Web service deployment to test the model further. After Eleni, Louis took over with ML topic on text corpus. The aim was to find optimal keywords to classify two diseases: neurodegenerative and heart disease. PubMed dataset was cleaned and pre-processed. Numeric feature vectors were extracted, and the classification and regression model was trained. Again, the model was scored, validated and deployed. Besides, Louis conducted a short tutorial on regular expression in the context of text mining.

The successful hands-on session was further enriched by Dimitrios 'tips-and-tricks' on how to apply for Azure and Horizon grants. The discussion about Azure grants was extended by other Azure grant holders, Petre Lameski and Eftim Zdravevski, who shared their experiences for making a successful application.

Topic III: Deep Learning

The last two days were guided by Tarry Singh, Founder/CEO of deepkapha.ai, with teaching and technical assistance of Raik Otto. They facilitated a workshop on Deep Learning (DL), which aims were:

[8]https://www.microsoftazurepass.com/

- to introduce the history and concept of DL;
- to grasp the power of artificial intelligence and what tremendous effect it could have on health care;
- to get participants acquainted with the use of Tensorflow and Keras on Python and its difference;
- to address DL practical issues with publicly available image datasets (MNIST and ISIC, respectively);
- to compare the performance with the different setups: using local CPU, local GPU (if available) and Microsoft Azure's NC6 GPU-backed instances.

Tarry Singh started with the history of ML and gave an overview of DL frameworks and libraries. He moved on to short introductions on several ongoing DL projects in healthcare and showed one active DL project from his company[9], under the hypothesis that (against all current statements) the thalamus is not just a passive relay centre. Furthermore, he exposed - amongst others - the pros and cons of the activation function ReLU (Rectified Linear Unit), which represents a significant step and is often used in DL frameworks. Tarry also stated that Aria2 (Adaptive Richard's Curve Weighted Activation) outperforms state-of-the-art activation functions on publicly available data such as MNIST, CIFAR10, and CIFAR100. He also mentioned the difficulties that computers have to differentiate between two similar looking things – and why humans don't have this problem as well as a possible solution to solve this problem with DL (Togootogtokh *et al.*, 2018).

Fortunately, for the most students support ticket was processed on time, meaning that the DL VMs were created and accessed at the beginning of the hands-on topic. Participants that were not able to create NC6 instances via their accounts were given access either by Eftim Zdravevski and Petre Lameski, from their Azure grant subscription or by Alberto Marcos González team, thus allowing the creation of additional NC6 instances to be used during the second session of DL.

After the theoretical introduction, the participants followed a hands-on tutorial about using Tensorflow to build DL models. The practical tutorial was performed using hands-on coding approach, where the participants were introduced to how to code and then were expected to finish or improve the task themselves.

In the first part, after the NC6 instance was created, the participants set up a secure connection to the machine with support for Jupyter Notebook so that the code could be executed in a more explanatory environment and not directly via the console. Tarry and Raik provided initial notebooks. They were cloned on the NC6 instances and then loaded from each participants computer using a local browser. The notebooks were partially filled with explanations and know-how, and with the helpful lead of Tarry and Raik, the blanks in the notebooks were filled by the participants.

First, the tutorial included how to load a dataset using Python and Tensorflow. For the first days exercise the participants were loading the MNIST dataset. The MNIST dataset consists of 60.000 training images of handwritten numbers and 10.000 testing images. After that, the layers of the neural network were added and configured (first convolutional layers and then the fully connected layer). For the convolutional layers the RELU activation function was used, and for the fully connected layer, we used the softmax activation function.

After the network was configured, the configuration of the training session was performed with code adding the type of optimisation and the other hyperparameters such as the number of epochs, dropout rate, learning rate etc. The participants were encouraged to tune the parameters of the network to obtain the best results on the MNIST dataset. This resulted in an unofficial competition which was won by Juliane Pfeil and Sabrina K. Schulze who managed to tune the parameters of the DL network and achieve an accuracy of 99.5 %. The participants were encouraged to try to improve their score as homework.

At the end of the day, the participants were able to load a dataset, design and configure a DL network, train a model and test its performance on the test set using Tensorflow and python.

On the last day of the training school, Tarry explained the difference between Tensorflow and Keras. Keras is a simple high-level neural networks library which works as a wrapper to Tensorflow, which makes DL programming easier and faster. To illustrate this, during the practical presentation, the same was performed with Keras, emphasising the shorter and easier to understand the approach that Keras uses. The training was once again performed for the MNIST dataset with remarkably little effort for the coding in comparison of using Tensorflow alone.

Then, the transfer learning approach was described and applied to train a DL neural network for skin cancer classification from mole image dataset (International Skin Imaging Collaboration - ISIC). Namely, the Inception V3 (Szegedy *et al.*, 2016) pre-trained weights were used, which were originally obtained when training on the ImageNet dataset. The ISIC dataset contains thousands of images from benign and malign skin cancer, and the task was to differentiate between them. All coding was performed together with Tarry, who explained all the steps and parameters on-the-fly. The participants were also introduced with the data generators capabilities that can load images from a folder, process them (resize, rotate, shift, etc.) to enrich the train or test batches. Another feature that was presented was the freezing and unfreezing of layers for training and adding early stopping criteria while training. All these options and some fine tuning of the network were used to improve the accuracy of cancer classification. Finally, the visualisation of the loss and the results was performed.

The last day was closed by a short closing remark from Tarry Singh and Marta Pérez-Casany.

[9] deepkapha.ai

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

Trainers' Perspectives

Eleni Papakonstantinou and Louis Papageorgiou

The ATS in Barcelona was an excellent opportunity for interacting with life science researchers and providing novel techniques and pipelines related to big data management and analysis through a well organised and successful workshop. Our aim through the 2-day session was to introduce the students to several of our projects in which cloud solutions have been applied and to implement a standardised pipeline for machine learning experiments using user-friendly and efficient platforms such as AMLS. We organised a hands-on session that all students could follow up regardless of their background, and experiment with different options for processing and analysing big data. Examples of predictive experiments were applied in two different directions, an analysis of the "Wisconsin Breast Cancer Diagnostics" dataset and a text learning approach for data mining. Even though there was a big setback with the Azure platform during the first day, students were able to discover the potential of cloud computing through AMLS in the next day and walk through a set of options provided to facilitate their research, regarding databases organisation, data processing, statistical analysis, feature extraction, machine learning algorithms and the deployment of a web service for predictive experiments.

Throughout the ATS we had the opportunity to interact with the other trainers and trainees, and have insightful discussions regarding highly important topics in our field of interest. In combination with the training session, which was a fruitful experience through the enthusiasm of the students, we consider this workshop as a valuable experience from which we can get important feedback on the real-life problems of both young and experienced researchers. We would also like to personally thank all the participants for contributing to the success of this productive and creative workshop, the fellow lecturers for setting their hearts and minds in organising the teaching sessions, and the trainees for their passion and enthusiasm. Many thanks to the organising committee (Ariel Duarte-López, Prof. Marta Pérez-Casany, and Prof. Erik Bongcam-Rudloff) for their warm welcome and all the efforts they made in front of and behind the scenes to accomplish this successful outcome. We strongly propose that this CHARME ATS. LS is a precursor for the progressive development of a unified scientific community able to address emerging scientific issues including the big data management.

Tarry Singh

The course and the logistics of the ATS were planned very well, and I wish to thank Ariel Duarte-López, Erik Bongcam-Rudloff, and especially my teaching assistant Raik Otto, who worked tirelessly to make this a successful workshop. It is the first time that we were able to do a full-scale DL bioinformatics workshop with real-world datasets. So, thanks not only to the organisers, but also the learners, especially Eftim Zdravevski for providing support as we struggled with the Microsoft Azure service support.

I was pleasantly surprised with the level of understanding and experience as well as enthusiasm most candidates displayed in understanding the concepts, conceptualising and reimagining into their projects and following the training diligently. It is always little time given there is so much to learn in this expanding field of DL, where new models, algorithms and network architectures are being updated on a regular basis. Having said that, it would be great to see a student/learner community emerge that pushes us to develop more advanced modules. In this way we can focus on solving more complex and challenging tasks, such as using best model with limited or smaller datasets, using Bayesian- or PGM (probability graph models) and eventually also experimenting with other advanced models such as GAN's (generative adversarial networks), Autoencoders, Capsule Networks and more.

My learning and take-away from this CHARME ATS was to collaborate with my peers (both professors, as well as students). There is a good scope of improvement when it comes to making a robust infrastructure ready for future exercises. I would suggest evaluating existing cloud computing as well as on-premise models and eventually choose for the right fit that allows learners to start immediately quickly. This also applies to trainers (myself included) that we must come with — for as much possible, a template-based approach so we can deploy systems before we start, and learners get on to learning from the first hour. Second, it should be good to consider planning a proper deep dive bioinformatics DL workshop where we spend at least three full days to learn and train the networks.

Apart from that, I am very grateful for all the generosity, flexibility and kindness that flowed throughout the ATS and I would be delighted to conduct more such workshops with you and perhaps even consider adjunct professor roles for helping build curriculums as there is a considerable need in medical institutions as well as the healthcare industry for such exercises. We keep getting bombarded by professors as heads of institutions who have a lot of data but lack skills and techniques to use DL.

Thanks once again and please stay connected as we are doing some excellent research as much as we are working with leading medical colleges and healthcare companies to make a more meaningful impact with DL and bioinformatics.

Miscellaneous

The first meetup was scheduled for Sunday, before the official start of the training school, where many participants as well as scientific committee and local organisers met and greeted at Palau de la Generalitat in the city centre of Barcelona to discover unique 'ir de

Hoang et al. (2019) EMBnet.journal 24, e917
http://dx.doi.org/10.14806/ej.24.0.917

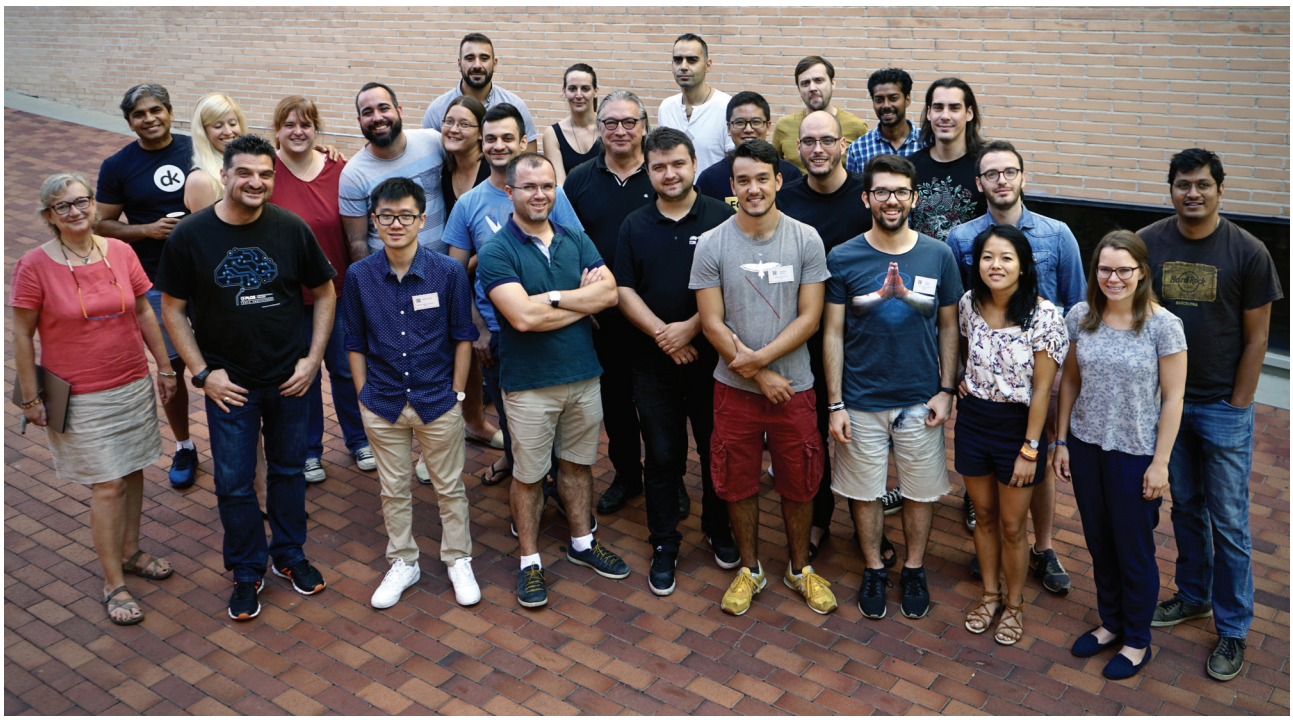

Figure 1. Trainees and trainers of the CHARME Advanced Training School on Big Data for Life Science (image provided by Sabrina Schulze).

tapeo' experience. The second meetup was scheduled on Wednesday after the hands-on sessions from Topic II to assure some 'intellectual break' and further enhance interpersonal relations. For this gathering, called 'potluck', all participants were encouraged to bring snacks, food or beverages from their countries to share. These social events arose from the suggestions of the Think Tank by the EMBnet network tradition. Many informal social events took place where participants, lecturers as well as assistants exchanged research experiences.

The success of this ATS can be measured to some extent by the average impression rates obtained through the participant evaluation forms. 'Content of course' scored on average 4.57 and 'course organisation' 4.86 (response rate of 14/23 and Likert scale [1,5]).

Conclusions

This paper described the topics, detailed content and timeline of events during the CHARME ATS on Big Data for Life Sciences, which took place at the Technical University of Catalonia (UPC) in Barcelona, September 3-7, 2018. All participants appreciated the supporting and inspiring environment in Barcelona. For this school, a more homogenous group was selected, and participants had the required level of prerequisite knowledge about the school's topics, thus streamlining the learning process. All participants seized the opportunity for scientific exchanges and discussions about ongoing projects. All lecturers motivated the participants to deep dive into this exciting topic of big data in bioinformatics. The training school was unanimously regarded as a well-organised platform to enhance future collaborations.

Abbreviations

AI	Artificial Intelligence
AMLS	Azure Machine Learning Studio
ATS	Advanced Training School
DL	Deep Learning
FS	Feature Selection
ML	Machine Learning
VM	Virtual Machines

Trainers' information

Speakers

Dr Andrzej Janusz, Assistant Professor at the Department of Mathematics Informatics and Mechanics, University of Warsaw, Warsaw, Poland (MIM UW), data scientist at eSensei[10], leading the SENSEI project co-funded by NCR&D and Co-founder of the Knowledge Pit platform[11]
Dr Dimitrios Vlachakis, Assistant Professor and group leader of the Genetics and Computational Biology group at the Department of Biotechnology, School of Food, Biotechnology, and Development at the Agricultural University of Athens in Greece
Dr Tarry Singh, CEO and founder of deepkapha.ai and AI neuroscience researcher, Amsterdam, Netherlands

Trainers' Assistants

Eleni Papakonstantinou, PhD (c), Department of Biotechnology, School of Food, Biotechnology and Development, Agricultural University of Athens, Athens, Greece

[10]http://esensei.net/

[11]https://knowledgepit.fedcsis.org/

Louis Papageorgiou, PhD (c), Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece, and Genetics and Computational Biology Group, Laboratory of Genetics, Department of Biotechnology, Agricultural University of Athens, 75 Iera Odos, 11855, Athens, Greece
Raik Otto, PhD (c), Institute for Mathematics and Computer Sciences, Humboldt University, Berlin, Germany

Declarations

Author's contributions

YH, JP, MZ, ŽR, AT, EZ, PL, TS, EP, LP contributed to writing the paper and conducted a critical revision. SKS contributed the notes. ADL and MPC proof-read and approved the final manuscript.

Acknowledgement

A special thanks to Domenica D'Elia, Erik Bongcam-Rudloff, Marcus Frohme and Kristina Gruden for their assurance and full support. A special note of recognition and gratitude goes to Marta Pérez-Casany and Ariel Duarte-López for leading the local organisation of the training school. Other thanks to all three lecturers and assistants who made this school absorbing and exciting. We want to thank the COST Action CHARME for providing us with the necessary funding for this event (COST Action CA15110 is supported by the EU framework program H2020). EZ and PL also acknowledge the support of Microsoft Azure for Research.

Availability of data and materials

Slides, examples, and exercises

FS:
bit.ly/2QHP2yA
https://github.com/janusza/
ML on Azure:
http://dimitrislab.com/azure/
DL:
bit.ly/2PNn6YS
https://github.com/RaikOtto/Barcelona
https://github.com/TarrySingh/

Publicly available datasets

Wisconsin breast cancer:
https://bit.ly/2pkdRnS
MNIST:
http://yann.lecun.com/exdb/mnist/
CIFAR:
https://www.cs.toronto.edu/~kriz/cifar.html
ISIC:
https://isic-archive.com/api/v1

References

1. Pfeil J, Schulze SK, Zdravevski E, Hoang Y (2018) Report on the "Big Data Training School for Life Sciences", 18-22 September 2017, Uppsala, Sweden. EMBnet.journal **23**, e905. http://dx.doi.org/10.14806/ej.23.0.905
2. Schulze SK, Ramšak Ž, Hoang Y, Zdravevski E, Pfeil J, Duarte-López A, Baier U, Zagorščak M. Proceedings of the "Think Tank Hackathon", Big Data Training School for Life Sciences Follow-up, Ljubljana 6th–7th February 2018. EMBnet.journal **24**, e912. http://dx.doi.org/10.14806/ej.24.0.912
3. Riza LS, Janusz A, Bergmeir C, Cornelis C, Herrera F, Ślezak D, Benítez JM. (2014) Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "roughsets". Information Sciences **287**:68-89. http://dx.doi.org/10.1016/j.ins.2014.07.029
4. R Core Team (2017). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (Accessed 03 September 2018)
5. Togootogtokh E, Amartuvshin A. (2018) Deep Learning Approach for Very Similar Objects Recognition Application on Chihuahua and Muffin Problem. arXiv preprint arXiv:1801.09573.
6. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. (2016) Rethinking the inception architecture for computer vision. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2818-2826).