EMBnet.journal
Bioinformatics in Action

Schulze et al. (2018) EMBnet.journal 24, e912
http://dx.doi.org/10.14806/ej.24.0.912

Reports

Proceedings of the "Think Tank Hackathon", Big Data Training School for Life Sciences Follow-up, Ljubljana 6th – 7th February 2018

Sabrina K. Schulze[1], Živa Ramšak[2], Yen Hoang[3], Eftim Zdravevski[4], Juliane Pfeil[5]✉, Ariel Duarte-López[6], Uwe Baier[7], Maja Zagorščak[2]

[1] Cell2Fab (Synthetic Biology, Faculty of Biochemistry and Biology), University of Potsdam, Potsdam, Germany

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

[3] Department of Signal Transduction, German Rheumatism Research Center, Berlin, Germany

[4] Department of Information Systems, Faculty of Computer Science and Engineering, Sts. Cyril and Methodius University in Skopje, Skopje, Macedonia

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

[6] DAMA-UPC, Department of Computer Architecture, Technical University of Catalonia, Barcelona, Spain

[7] Institute of Theoretical Computer Science, Institute of Theoretical Computer Science, Ulm University, Ulm, Germany

Competing interests: SKS none; ŽR none; YH none; EZ none; JP none; ADL none; UB none; MZ none

Abstract

On 6th and 7th February 2018 a Think Tank took place in Ljubljana, Slovenia. It was a follow-up of the "Big Data Training School for Life Sciences" held in Uppsala, Sweden, in September 2017. The focus was on identifying topics of interest and optimising the programme for a forthcoming "Advanced" Big Data Training School for Life Science, that we hope is again supported by the COST Action CHARME (Harmonising standardisation strategies to increase efficiency and competitiveness of European life-science research - CA15110). The Think Tank aimed to go into details of several topics that were - to a degree - covered by the former training school. Likewise, discussions embraced the recent experience of the attendees in light of the new knowledge obtained by the first edition of the training school and how it comes from the perspective of their current and upcoming work. The 2018 training school should strive for and further facilitate optimised applications of Big Data technologies in life sciences. The attendees of this hackathon entirely organised this workshop.

Background and course

In September 2017 the first "Big Data Training School for Life Sciences"[1] as organised by the COST Action CHARME[2] and EMBnet[3] and took place in Uppsala, Sweden (Pfeil *et al.*, 2018a). Although the academic background of young bioinformaticians participating in this training school was somewhat diverse, we all concluded that it is a necessity to learn more about handling Big Data in life sciences. It was denoted how it would be a great deal if additional useful topics could have been described, with ones already presented in Uppsala but shown in more detail. Out of these constructive discussions the idea of additional training came up, and it was proposed to the members of CHARME and EMBnet. This idea was welcomed with interest and approved by the organisers in the form of a short pre-meeting, which was organised by us. As a result, the Think Tank Hackathon[4,5] was organised at the Medical Faculty of the University of Ljubljana, Slovenia on 6th and 7th February 2018 with the support of ELIXIR.SI Node[6].

The hackathon was only open to attendees of the original training school in Uppsala, with eight of them enthusiastically participating (Figure 1). The organisation of this event was made under the leadership of Maja Zagorščak and Živa Ramšak, with the extra help of all the other attendees. As all the participants share the trait *"don't just tell me how to do it, I want to figure it out really"*,

[1] http://astrocyte.com/COST-CHARME/COST-CHARME/Home.html

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

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

[4] http://conferences.nib.si/BigData/

[5] http://www.cost-charme.eu/events/follow-up-training-school

[6] https://www.elixir-europe.org/about-us/who-we-are/nodes/slovenia

Article history
Received: 19 February 2018
Published: 19 April 2018

Schulze et al. (2018) EMBnet.journal 24, e912
http://dx.doi.org/10.14806/ej.24.0.912

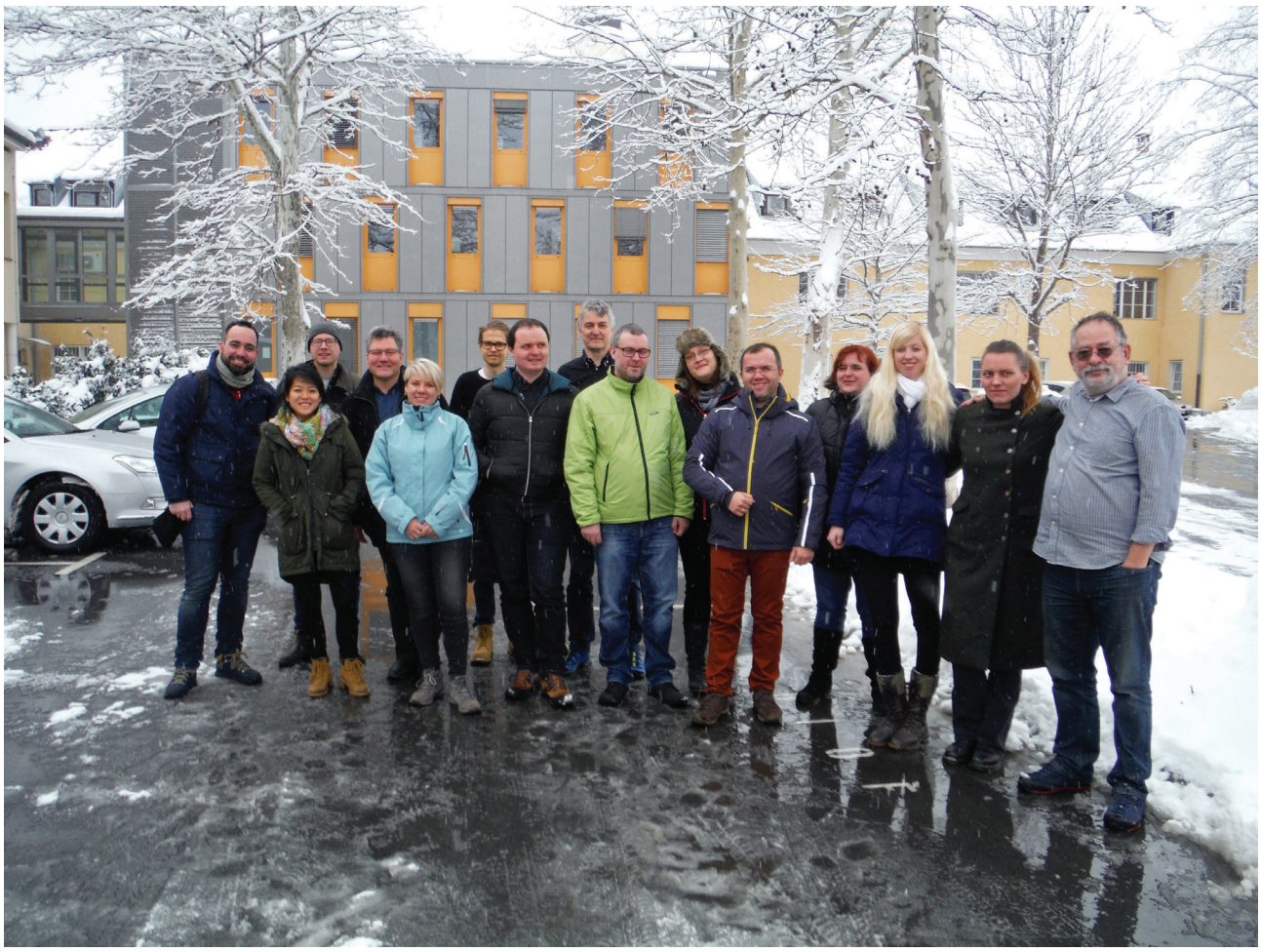

Figure 1. Attendees of the Hackathon joined by COST Action CHARME and ELIXIR.SI representatives (image kindly provided by Sabrina Schulze).

the primary focus of the meeting was the exchange of substantial knowledge about leading technologies and *in situ* troubleshooting by solving some real problems.

Before the training school, with the help of a questionnaire regarding group expertise and interests, and taking advantage of various collaboration technologies (*e.g.*, Slack, GitHub and Dropbox), we identified and agreed on several feasible topics, such as *Code, Parallelize, Containerise!, Image analysis with OpenCV - from leaf shape to chords, Machine learning - do it yourself* and *Deep learning without a PhD*.

The attendees, each knowledgeable in their subfield, took the complementary lead on diverse topics and were solely responsible for preparing presentation materials, code examples and tasks for the hackathon. During the Think Tank in Ljubljana, all issues were covered by the active participation of all attendees. Additionally, we brainstormed (hence the name Think Tank) and all together collaborated to the drafting of a proposal for a potential advanced training school.

The first day was opened by Maja Zagorščak with some introductory words about the Think Tank Hackathon goals, organisation and schedule. Her introduction was followed by Eftim Zdravevski, on the topic of Docker technology and its Big Data applications. He discussed parallelisation using Apache Spark™ (Zaharia *et al.*, 2010) and demonstrated k-means clustering and classification examples using the MLLIB (Meng *et al.*, 2016), Spark's scalable machine learning library. Therefore, he guided us through the process of algorithm parallelisation using the data-parallel computing paradigm, while discussing its performance and computational acceleration, using an approach similar to the ones presented Zdravevski et al., 2015a and in Zdravevski et al., 2015b.

Juliane Pfeil filled the second part of the day with a lecture on OpenCV, a computer vision library. She guided us through basic concepts of image analysis (*i.e.*, noise reduction, segmentation, structural pattern recognition and prediction) and demonstrated which functions and parameters are optimal for specific predetermined tasks (for an example see Pfeil *et al.*, 2018b). Lastly, she showed an interesting example of how detection of the centre of mass and radial function application can be used to transform bioimages into chords. All source code developed during the Think Tank hackathon is available online at the GitHub repository[7] Big Data Think Tank.

The second day began with *Deep learning without a PhD*, where Yen Hoang introduced Google's TensorFlow open-source library, which among others is also used for

[7]https://github.com/zagorGit/BigDataThinkTank

Schulze et al. (2018) EMBnet.journal 24, e912
http://dx.doi.org/10.14806/ej.24.0.912

Figure 2. The attendees during the deep learning session (image kindly provided by Sabrina Schulze).

machine learning applications (*e.g.*, neural networks). She guided us through various detailed presentations by Martin Görner (GitHub repository: TensorFlow MNIST Tutorial[8]) and moderated the discussion about neural network theory and the underlying code to increase our comprehension. For this section, the entire infrastructure was set up in advance by the ELIXIR.SI's member Andrej Kaštrin (users, applications, dependencies), which resulted in no time loss for the participants once the workshop started. Additionally, this led to an informal discussion over Docker usability in R, to ease issues with transparent data sourcing, availability and traceability of published results. These concerns are also of vital importance considering principles of FAIR (findable, accessible, interoperable, reusable) scientific data management (Wilkinson *et al.*, 2016). The presentations were concluded by an impromptu presentation given by Uwe Baier, regarding his ongoing work in data compression with examples. Among them the magic of the Burrows–Wheeler transform (BWT; Baier *et. al.*, 2016). Aside from standard BWT application in text compression, it was shown advantageous in many next-generation sequencing (NGS) alignment algorithms in the context of memory reduction.

[8]https://github.com/martin-gorner/tensorflow-mnist-tutorial

Both days were rounded off by discussions and brainstorming sessions regarding the organisation and structure of a potential follow-up training school to be organised later in 2018. We discussed about a possible organising committee, suitable dates and location of the training school; all these details will be formulated into an official proposal that will be submitted to CHARME for approval.

The "Advanced" Big Data Training School for Life Sciences was proposed to last 5 days, with various topics suggested to be focused on: feature extraction, deep learning, with both free of charge and payable software tools and platforms (*e.g.*, services offered by Microsoft Azure), or the backbone computational infrastructure required for tasks in Big Data analysis. All of these topics represent attractive options, given the current developments in these fields, especially if the knowledge learned is connected to real-world problems. It was also agreed to invite experts on these topics, which would explain the theory behind these applications and assist during the practical parts of the workshop. These ideas were finalised on the evening of the second Think Tank day, where preliminary time slots were agreed upon, including time for theory sessions and practical parts. The majority of practical sessions were proposed to be group tasks, where the participants would jointly work on some pre-selected datasets. The results of practical sessions

would then be presented to the remaining groups, thus optimising the time constraints of such tasks, and at the same time allowing for a continual learning improvement via the collaboration of the workshop's attendants. On that note, social activities were also taken into account in the schedule, again fostering future cooperation of participants. We finally decided that participants of the previous training school should be preferred as attendees for the advanced training school. The remaining slots should then be filled by other applicants, provided they are eligible given their background.

At the end of the Think Tank, a biweekly journal club was proposed, where related papers will be read beforehand and discussed in a one-hour session. In the beginning, this should be held only with the participants at the Uppsala training school. After some routine, collaborators could join the video conference via Skype or similar. This club would allow a progressive alignment of the participants' level regarding machine learning and deep learning elements.

Summary

This Think Tank event turned out in two exciting and productive days. We established the knowledge base on four different topics and identified topics for a possible advanced Big Data training school. The attendees who gave presentations on the hackathon noticed how it is sometimes easier to suggest improvements than to apply them in practice. However, the intellectual freedom in the organisation of future events gave us the opportunity to think outside the box. We consider the results of these days as a useful starting point leading towards a proposal for the Advanced Training School. Additionally, the idea of a biweekly journal club would help to bring all participants on the same level and to strengthen the network.

Fortunately, there was also some time scheduled for social events and networking. One was on the evening before the first session, where we spent some time playing abstract games in a board-game tavern, thus getting acquainted with other participants mindsets in a relaxed and casual manner. This idea was well accepted. Therefore, we suggest that this type of social activity is also included in other follow-up events. It offers a unique opportunity of catching up or getting to know potential new participants better and in advance to the starting of the official training event. The other social event we attended was on the evening of the first day. It was the dinner with CHARME WG5's meeting attendants. Some of them were in Uppsala the last September and to meet them in Ljubljana helped us to set up foundations for future collaborations and ideas on how to get the wheels in motion with the organisation of the "Advanced Training School".

Acknowledgement

We would like to thank ELIXIR.si for allocating rooms for us at the University of Ljubljana (Faculty of Medicine) and for infrastructural support. We also thank the COST Action CHARME (CA15110) that is supported by the EU framework program H2020, for providing us with the necessary funding for to organise and attend this event. A particular note of acknowledgement and gratitude goes to Maja Zagorščak and Živa Ramšak for preparing the collaboration tools and environments, for to schedule and leading the organisation of the event in the highest of standards.

References

1. Baier U, Beller T, Ohlebusch E (2016) Graphical pan-genome analysis with compressed suffix trees and the Burrows–Wheeler transform, Bioinformatics **32**, 497–504. http://dx.doi.org/10.1093/bioinformatics/btv603
2. Meng X, Bradley J, Yavuz B, Sparks E, Venkataraman S *et al.* (2016) Mllib: Machine learning in apache spark. J Mach Learn Res **17**, 1–7. http://jmlr.org/papers/v17/15-237.html
3. Pfeil J, Schulze SK, Zdravevski E, Hoang Y (2018a) 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
4. Pfeil J, Frohme M, Schulze K (2018b). Mobile Microscopy and Automated Image Analysis: The ease of cell counting and classification. Optik & Photonik **13**, 36-39. http://dx.doi.org/10.1002/opph.201800002
5. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**. http://dx.doi.org/10.1038/sdata.2016.18
6. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: Cluster computing with working sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud'10). http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-53.html
7. Zdravevski E, Lameski P, Kulakov A, Jakimovski B, Filiposka S and Trajanov D (2015a) Feature ranking based on information gain for large classification problems with mapreduce. IEEE, Proceedings of the Trustcom/BigDataSE/ISPA. http://dx.doi.org/10.1109/Trustcom.2015.580
8. Zdravevski E, Lameski P, Kulakov A, Filiposka S, Trajanov D and Jakimovski B (2015b) Parallel computation of information gain using Hadoop and MapReduce. IEEE, Proceedings of the FedCSIS. http://dx.doi.org/10.15439/2015F89