Towards AI services à la carte: Status and Future Work at University of Oslo

Anne Schad Bergsaker [1], Andrea Gasparini [123] and Thomas Röblitz [1]

[1] University Center for Information Technology, University of Oslo, Oslo, Norway
[2] University of Oslo Library, Oslo, Norway
[3] Department of Informatics, University of Oslo, Oslo, Norway
a.s.bergsaker@usit.uio.no
andreg@ifi.uio.no
thomas.roblitz@usit.uio.no

Abstract. The accelerating digitalization of nearly every aspect in society leads to unprecedented amounts of data produced every day. While this provides a great opportunity to analyze for the benefit of society it also poses several challenges among others (A) legal, (B) ethical, (C) modeling complexity, (D) shortage of processing power, (E) shortage of data storage, (F) many users being non-IT experts. Employing artificial intelligence (AI) methods provides a promising approach to tackle challenge (C). The advent of capable computing hardware (GPUs) and easy-to-use programming frameworks have dramatically lowered the entrance into data science for researchers, businesses and the public sector and thereby help addressing challenges (D) and (F). In an explorative effort, two service centers at the University of Oslo have partnered to venture into the field of AI to develop a portfolio of services dubbed "AI services à la carte". This service portfolio shall enable our users to make good use of AI methods in an efficient, secure and unbiased way leading to excellent education and research. We present the status and future of our work towards such a service portfolio.

Keywords: Artificial Intelligence, Research Infrastructure, Education, Services

1 Introduction

Fueled by digitalization which provides unprecedented amounts of data every day science has developed into "The Fourth Paradigm: Data-Intensive Scientific Discovery" [1]. The first step in analyzing vast volumes of data was focused on scaling existing algorithms to extreme levels. This led to innovations such as MapReduce [2], DynamoDB [3] and many more which are building blocks of Clouds [4] as we know them today. However, all these innovations alone could not overcome the fundamental bottleneck of "human intelligence" to develop sophisticated models and to implement these in the prevailing paradigm computational science. Already for a long time, machine learning (ML) methods such as linear regression, support vector machines, k-means clustering, etc. provide well-understood means to perform classical statistical data analysis. Recently, deep learning (DL) has proven to yield higher accuracy for some applications such as speech recognition, object detection and language translation. The advent of capable computing hardware (GPUs) and easy-to-use programming frameworks such as TensorFlow, PyTorch and Keras made deep learning and traditional machine learning methods accessible to a wide audience of data scientists. To fully exploit this new powerful tool in a university environment several areas require significant and coordinated work:

- **Technological basis:** modern resources need to be offered in a flexibly way to satisfy diverse needs of researchers and students.
- **Legal framework:** users need to understand legal limitations of their work to avoid any pitfalls when working with large datasets.
- **Ethical considerations:** users need to be aware of ethical aspects such as bias in data which could lead to discriminating models.
- **Courses:** all stakeholders will require a significant buildup of competence in several aspects.
- **Community building:** historically, the adoption of AI methods by enthusiasts has led to significant fragmentation in the community.
- **Competent service staff:** to consult users in choosing the services they need service staff must build up knowledge in several fields.

Eventually we expect that each area is matched with a set of services which all together build the portfolio (or “menu”) of the “AI services à la carte”. At University of Oslo (UiO) two service centers – the Oslo University Library (UBO) and the University Center for Information Technology (USIT) – have started joint activities to prototype specific areas of that service portfolio. We have chosen to adopt an agile approach (fail early, learn from user feedback) and engage in several small-scale activities with partners at the university as well as external organizations to incrementally build up that service portfolio. By no means the collaboration of UBO and USIT is meant as an exclusive club, it just happened that both institutions found common ground from early on. We see similar activities in Norway as well as worldwide, for example, at MIT [5].

In the remainder of the paper (Section 2 to 5) we describe activities we pursued for a subset of the areas listed above. The work we report has started about 15 months ago, hence it is still in an early phase. Nevertheless, we think the work is of interest to other universities or research organizations, and we are very interested in seeking collaborations to exchange knowledge or even work on some of the areas together. We conclude the paper with summarizing major findings and directions for future work in Section 6.

2 Technological basis

One pillar of our work is to set up a modern technical infrastructure tailored at workloads specific to machine learning, deep learning and data science in research and education. As result of numerous discussions with scientists and lecturers we identified three basic building blocks for hardware: (1) accelerators such as GPUs, (2) large main memory (RAM) and (3) fast storage (for example SSDs). Most users seem to only require high-level frameworks such as TensorFlow, PyTorch, and Keras as the basic software environment which is readily available via the popular Anaconda package management system.

For procuring a first setup we evaluated various options based on the following criteria: training/inferencing performance, performance to cost ratio, support for popular programming frameworks, ability to support different provisioning means (physical servers or virtual machines), hardware variety to gain experience, and needs not covered well by the current e-Infrastructure in Norway. Our initial experimental setup includes five servers each equipped with 128 GiB RAM, 200 GB fast storage (SSD), 6 TB bulk storage (HDD), 28-32 Intel/AMD CPU cores and different GPU options: 3 servers with 4 x NVIDIA RTX 2080 Ti, 1 server with 1 NVIDIA T4 and 1 server with 2 x AMD Vega 64. The RTX and Vega servers are available for researchers and are also used by students in a class. The T4 server is being integrated into the Cloud-like UH IaaS [6] and we hope to use this to experiment with virtual machines supporting virtual GPUs.

This setup was found to be a good startup and quickly became popular among scientists and students. We found that consumer grade GPUs (RTX 2080 Ti) provide an excellent performance/price ratio but require special attention to ensure stability. Mixing workloads (researchers and teaching classes) on stand-alone physical servers lowers the overall utilization. Using servers' local bulk storage to construct a shared file system is easy to setup, e.g., with BeeOND [7], but affects overall system availability if servers need to be rebooted due to instable operating of consumer grade GPUs. Local fast storage of approximately 200 GiB for both OS and data is too small.

In the short-term future we will develop the infrastructure further to address the shortcomings observed so far, particularly, we will increase local fast storage, separate bulk storage from compute servers and enable separation of workloads by hosting teaching classes via the JupyterHub framework [8] on the UH IaaS. We may also investigate putting some form of workload management system into place to let users share the scarce resources in a fair and predictable manner.

3 Courses

Making good use of AI methods in research and education requires substantial efforts in building up knowledge by researchers, IT staff and lecturers. We envision that a research organization such as University of Oslo shall establish a comprehensive list of courses addressing users with different level of expertise, cover data management and preparation methods, teach technical aspects to utilize existing resources efficiently, include non-technical topics such as legal and ethical considerations, and last but not least provide a solid foundation of the underlying theory in machine learning, deep learning and data science. To explore this vast field, we have conducted several activities and collaborate with existing initiatives such as the Carpentries at University of Oslo [9].

One activity was to organize a full-day workshop for this year's Research Bazaar [10]. Using a large papyrus dataset from a previous AI project at the library [11], the workshop emphasized the need for proper data organization to successfully apply machine learning. Participants were first introduced to the dataset, then they learned about Design Thinking [12] and applied this to prototype new approaches to analyze papyruses. Finally, three ideas were implemented in Jupyter notebooks using basic techniques in Natural Language Processing and machine learning. Using Design Thinking helped participants to not start coding right away but redefine the problem to be solved first.

Another activity is trying to enable the use of AI technology available through large Cloud providers by hosting workshops held by external experts. So far, we had one workshop with Microsoft, where participants were able to train their own models using the Azure framework which provides a very simple interface without the need for traditional programming. It proved to be a very good starting place for researchers and students who have limited experience with programming and computational science. One caveat we found with using AI technology on external Clouds is that training of large models can get expensive over time. Nevertheless, our intention is to continue with such workshops, demonstrating other Cloud-based solutions as well, such as Google Cloud Platform.

A third activity started in January 2019, where we became involved in a bachelor and master course at the Law faculty concerned with the use of AI in different contexts of law practice. We provide assistance to add lectures fitting the needs and background of the law students, many of which have never programmed. We conduct a half-day workshop on Design Thinking which enables the students to prototype ideas without the need to know anything about programming. Thereafter, Microsoft will give a workshop on their Azure platform. The experience with that activity is that lecturers benefit from having a single place where they can get comprehensive advice addressing all the different perspectives of using AI in education.

By running activities as the ones outlined above, we want to establish a portfolio of course components which can be flexibly combined and offered based on needs of researchers or lecturers. We will collaborate with the Carpentries to construct a set of basic practical lectures to enable scientists an easy uptake of AI technology.

4 Community building

Since the 1950s, AI and machine learning have been a topic in research with several phases of high activity followed by low progress. Typically, research groups owned their own hardware dedicated at machine learning tasks. IT centers did not provide any common infrastructure base nor had significant competence to assist potentially interested researchers. While we recently witnessed a surge in interest and researchers actively starting to employ AI methods in their work, the fragmentation in the community continues to exist to a large degree. To help building a lively community, we begun in the fall 2018 to organize monthly events, named *AI lounge* [13], where researchers, students, IT staff and industry meet to share ideas, present their work, demonstrate solutions, and initiate new collaborations. The AI lounge establishes a low entrance barrier to a forum where essentially everyone interested in the field can contribute to, has attracted up to 60 participants, and already lead to new collaborations. It also gives us a unique possibility to map out the landscape of needs, activities, competence at the University of Oslo. Furthermore, the goal for this activity is to create an arena for sharing competence and convergence between relevant actors. The latter has happened in several occasions, not only between industry and researchers, but also inside the organization as several groups in the university are actively engaging in using AI technology in their day to day activities.

5 Service staff competence

The need to improve competence in using AI technology amongst researchers and lectures is mainly addressed by activities described in Section 4. To be able to effectively help our users we need to improve our own competence in the field significantly. In this section we describe several activities in which we seek to implement meaningful IT services exploiting AI technology thereby building up our own expertise.

One activity aims at supporting research by developing a service that can recognize dig sites in photographs from the natural history and cultural history museums in Oslo. The museums possess a large database of photographs which currently are tagged manually. Typically, the images are either of specific type of items, animals or dig sites. Besides tagging dig sites, the service shall also be able to filter photographs which do not contain dig sites. Over time the goal is to train more advanced models with more categories. That way, when new photos are added to the database, they may not need to be manually tagged by the user but can be automatically tagged by the model.

A second activity that has just started, is a collaboration between USIT, the text-lab at UiO, and the Norwegian national broadcasting corporation (NRK), in which our goal is to train a model that can recognize Norwegian speech and transcribe it automatically. As many of our researchers work with interview data, transcription is a process that takes up much time. In addition, many interviews include sensitive data, hence researches are not able to use external cloud-based services providing readymade trained models. While NRK may not be restricted by data protection laws, transcribing their vast archives on commercial clouds could become very expensive.

We plan to start more such “in-house” activities, for example, to analyze vast logs of system health monitoring data to better understand the state of IT infrastructure we operate. Also, we will seek collaborations with experts on areas such as law and ethics which are important topics in AI, but they usually do not fall into our centers’ responsibility. As we build up our own competence, we may also engage in user projects to help them build new services or apply AI technology in demanding research activities.

6 Major findings and future directions

Our major findings and plans for future directions for the presented areas are:

Technical basis. Our initial procurements proved to be well chosen and by maintaining close interactions with researchers and lecturers we identified future updates and upgrades: more local fast disk, separate bulk storage, and distinct platforms for research and teaching workloads.

Courses. We have conducted several courses, workshops and contributed to lectures addressing different topics and audiences. We will continue this activity as it is vital to help educating next generation data scientists in employing AI.

Community building. Bringing people together, e.g., with the AI lounge, was received very well, and we are looking forward to continuing this activity by inviting internal and external speakers on any topic related to using AI.

Service staff competence. Apart from building up competence through activities in all the other areas, we have started small-scale projects to use AI in "in-house" services. We plan to start more of such projects and continue seeking opportunities to participate in research projects.

In addition to these four areas, we will seek to complement the service portfolio with currently unaddressed areas "Legal framework" and "Ethical considerations" by partnering with experts internal or external to UiO. Through our work we may also discover new areas where users need advice from competent service staff, for example, support for developing of AI-based services and methods, hosting AI-based services for research and education, copyright of results achieved with AI, trustworthiness of model results, implications of the EU GDPR [14], and educational perspectives such as pedagogy, gender, ethnicity, equal rights and so on.

Within the next 12 months, we want to move from the current explorative phase towards production by establishing a starting set of basic high-quality services.

Finally, we seek exchange and collaboration with research organizations external to UiO to continuously improve our service portfolio.

Acknowledgments. The authors wish to thank all workshops participants for the insight they made available. We thank the Council for E-infrastructure (Fagråd for eInfrastruktur) at University of Oslo for supporting our work with funding for hardware, workshop hosting and staff. Finally, we acknowledge the crucial role of National Library of Norway that partly financed our activities (Research Bazaar 2019).

References

1. Hey, T., Tansley, S. & Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research (2009).
2. Dean, J. & Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, 137–150. San Francisco (2004).
3. DeCandia, G. et al.: Dynamo: Amazon's Highly Available Key-Value Store. Proceedings of the 21st ACM Symposium on Operating Systems Principles, Stevenson, WA, (2007).
4. Fox, Armando, et al. Above the clouds: A berkeley view of cloud computing. Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS 28.13 (2009).
5. MIT reshapes itself to shape the future, http://news.mit.edu/2018/mit-reshapes-itself-stephen-schwarzman-college-of-computing-1015, last accessed 2019/4/10.
6. UH IaaS, http://www.uh-iaas.no, last accessed 2019/4/17.

7. BeeOND – BeeGFS On Demand, https://www.beegfs.io/wiki/BeeOND, last accessed 2019/4/17.
8. JupyterHub, https://jupyterhub.readthedocs.io/en/stable/, last accessed 2019/4/17.
9. Carpentries at UiO, http://www.ub.uio.no/english/courses-events/courses/other/Carpentry/, last accessed 2019/4/10.
10. Hands-on Workshop: Exploring Research Data with Artificial Intelligence and Design Thinking, https://www.ub.uio.no/english/courses-events/events/all-libraries/2019/research-bazaar/190111_ResearchDataAI.html, last accessed 2019/4/10.
11. Gasparini, A., Mohammed, A. A., & Oropallo, G.: Service Design for Artificial Intelligence. ServDes.2018 Conference Proceedings Co-Creating Services, 1064–1073. Milano (2018).
12. Brown, T. Change by design: how design thinking transforms organizations and inspires innovation. New York: HarperCollins Publishers (2009).
13. AI lounge at UiO, https://www.uio.no/tjenester/it/forskning/kompetansehuber/uio-ai-hub-node-project/events/, last accessed 2019/4/17.
14. The EU General Data Protection Regulation, http://eugdpr.org/, last accessed 2019/4/10.