Collaborative Data Science Projects

Fourth Call for Proposals 2020

Data Science is one of the Strategic Focus Areas defined in the Strategic Planning 2017–2024 for the ETH Domain. The ETH Domain has launched the Initiative for Data Science in Switzerland (IDSS) to foster the adoption of data science through education, research and the provision of infrastructure. The Initiative creates the Swiss Data Science Center (hereafter SDSC), whose mission is to accelerate the use of data science and machine learning techniques within academic disciplines of the ETH Domain, the Swiss academic community at large, and the industrial sector.

So far, about two thirds of the funding from the ETH Board have been allocated to 39 Collaborative Data Science Projects that were reviewed and selected on a competitive basis. This budget has been used to cover (i) the costs of the SDSC's staff members assigned to the projects, (ii) the operational expenses and, in particular, the provisioning of compute and storage resources at SWITCH and CSCS, and (iii) the costs of the scientific and technical staff specialized in the scientific domain of research and employed by the applicant's institutions.

Similar to previous calls, the goal of this fourth call for SDSC Collaborative Data Science Projects is to offer the possibility for ETH Domain laboratories to benefit from the expertise of the SDSC and to undertake joint projects with a strong data science component. Each Collaborative Data Science Project is proposed and led by one main applicant from an ETH Domain institution.

A. Presentation of SDSC Collaborative Data Science Projects

1. General goals

The following observations motivate the SDSC call for Collaborative Data Science Projects:

- For many domain experts working with data, it is difficult to benefit from the most recent techniques in data science. Many approaches, including among others machine learning and deep learning techniques, are often hard to exploit in real world applications, given their complexity and fast paced progress in the last decades.
- For researchers working on theoretical aspects of data science, applying their newly developed techniques to interesting problems requires to put in place sizable collaborative projects, entailing endeavors that are difficult to undertake and sustain alone.

The goals of the SDSC Collaborative Data Science Projects are thus to help researchers and domain experts leverage the state-of-the-art in data science to develop models and analyses to support their research. At the same time, SDSC collaborative schemes aim at supporting the application of techniques developed in research labs working on data science methods to real world scenarios.

The scope of Collaborative Data Science Projects is representative of the diversity of the research undertaken within the ETH domain. A number of these research projects are in environmental sciences, in health and biomedical sciences as well as in the physical sciences, and are thus well aligned with current strategic research directions defined by the ETH board. The SDSC collaborates in other domains such as economics or social sciences as well.

In most projects, it is key to build models that are likely to produce scientific insights, which thus have to be interpretable and to integrate domain science prior knowledge. In a number of applications, they need to propagate and predict accurately uncertainties, often in complex structured models, and to be robust to non-stationarities. Data availability is a condition to get started in most SDSC collaborative projects, although the latter are not limited to modeling and analyzing data but can also focus on the optimization of data acquisition processes or on the interactions between scientists and algorithms. More information about ongoing SDSC Collaborative Data Science Projects can be found at https://datascience.ch/academic-projects/.

2. Expertise provided by the SDSC

Through Collaborative Data Science Projects, the data science team at the SDSC makes available to partnering ETH Domain research teams expertise which broadly covers data science modeling techniques including statistics, statistical signal processing, machine learning, deep learning, and techniques from computer vision, natural language processing and optimization, as well as computational methods for these techniques. Many of our data scientists also have expertise in specific domains and their corresponding data science approaches. In particular, we hope to not simply help in the choice of the standard models and techniques required to solve a given problem, but rather to build models tailored to the application considered, and that best incorporate the prior domain expertise into state-of-the-art data science techniques

3. Expected mode of collaboration

Successful projects in data science often crucially rely on strong exchange between domain experts and data scientists. The SDSC is fully invested in the projects in which it collaborates and typically leads or co-leads several work packages. We expect the domain expert collaborators to be also actively involved in the project. Their expertise is to guide the development of the methods and participate directly in data exploration and analyses.

In most of our projects, the SDSC data scientists working on the project meet at least every other week with collaborators from the partnering laboratories. These meetings can be at the SDSC, in the partnering lab or via teleconference. Given the important role of communication and informal exchange of ideas in interdisciplinary research, for some projects, our data scientist can regularly spend time in one of the partnering labs. This question is examined on a case by case basis. We also organize bi-monthly progress meetings with all collaborators, PIs and team leaders from SDSC.

4. Expected outcomes of the projects

The outcomes of collaborative projects with the SDSC are typically:

- scientific papers published in the journals of the community of the domain scientists;

- communications in conferences in the domain science;
- scientific code and data libraries that implement the methods and algorithms; and,
- possibly, papers in data science journals or conferences, if the project required specific advances in data science methods.

5. About open science and FAIR principles

The SDSC strongly supports the development of open science and FAIR (findable, accessible, interoperable, reusable) research practices. The SDSC engineering team developed Renku, a platform which facilitates traceability, reproducibility, and collaboration in data science projects. The Renku platform makes it possible to structure and manipulate complex datasets, together with their metadata and to export them easily to platforms assigning a unique DOI such as Zenodo and Dataverse. At the same time, Renku offers an integrated tool to track dependencies between code, data, and outputs which ensures full reproducibility at any given time. More details can be found at https://datascience.ch/renku/.

B. Fourth Call conditions, structure and characteristics

The project proposals will be evaluated both by external reviewers (researchers with specific domain expertise and, if possible, data science expertise) and by internal reviewers (SDSC senior data scientists and faculty members on the SDSC board).

The evaluation of each project will be based on an assessment of: its scientific quality, its impact from the domain perspective, the relevance of leveraging/designing specific data science techniques and the relevance of the SDSC expertise. Of course, the feasibility and risk assessments of the work packages of the project will be considered, in particular given the availability of data at different stages of the project, the availability of the appropriate domain expertise(s), and the complexity of the objectives involved. At the pre-proposal stage, a general description of a specific data science approach is not expected, and an explanation of the potential relevance of data science techniques to the considered problem suffices. A full-fledged description of the proposed data science approach will only be required in the full proposal. For the data science approach, applicants of selected pre-proposals will be invited to exchange with data scientists from the SDSC end of August/ beginning of September to elaborate or refine it in view of the full proposal.

The call will follow the calendar below:

Call for pre-proposals:	June 2, 2020
Submission of pre-proposals	July 13, 2020
Invitation to submit full proposals	August 10, 2020
Submission of full proposals	October 26, 2020
Final decision on full proposals	January 25, 2021
Start of projects	March-June, 2021

The remainder of the document describes the general conditions for Collaborative Data Science Projects, funded resources and a general description of the submission material requested.

1. General conditions

- Only researchers employed by an institution of the ETH Domain are eligible as Principal Investigator (PI), namely EAWAG, EMPA, EPFL, ETHZ, PSI, and WSL.
- Principal Investigators must independently direct the research and manage the resources they receive for the duration of the project as well as coordinate the efforts among the different partners. These are typically Professors and Senior Collaborative Scientists. We also accept Postdocs but since they are non-permanent staff, their application must be backed by a Professor (head of the lab), as co-PI.
- The project duration is limited to 24 months.
- The expected budget of a project is between 200'000 and 600'000 CHF.
- Pre- and full proposals are to be submitted in English.
- The application process has two-stages: a first step during which pre-proposals are submitted, and a second step, in which selected applicants are invited to submit a full proposal.
- Proposals that are continuation of funded and undertaken projects from a previous call can be submitted. They will be considered as new proposals and be subject to the same review criteria as new proposals.
- The PIs of funded projects will be requested to produce every year written scientific and financial reports that will be reviewed by the SDSC steering committee. Scientific reports detail the progress made during the past year, activities foreseen for the next one and any changes or deviations from the initial project plan. Financial reports should detail all spending during the period.
- The Management Office of the SDSC will assist in the administration and coordination of the project and will ensure accountability of the project according to the directives defined.

2. Funded resources

a. SDSC Technical Staff

The funding request personnel from the SDSC involved in collaborative projects is typically at least 50% of an FTE of a Data Scientist for 24 months.

b. Staff hired in one of the partnering teams

If necessary for the project, the applicant can request funding to hire personnel at the postdoc or PhD level contributing domain expertise (and usually with an interest in data science) for a duration of at most 24 months. It is also possible to request funding for a part-time research or computer engineer (e.g. at 10-25%). It is expected that the partnering laboratories contribute in kind to the project (see the budget proposal template for details).

c. Compute and Storage Resources

The SDSC uses resources from CSCS, SWITCH, or the computational resources already available to its collaborators. The cost of the computing resources (CSCS, SWITCH) needed by the project is covered in the funding of the project. More precisely, the following resources can be requested:

- CPU Cores and RAM (through virtual machines or compute time on HPC clusters);
- Data storage; and,
- GPU boards.

They have to be budgeted in the annex of the (pre-)proposal. The acquisition of new hardware (e.g. personal computers) is not covered, nor is the computing and storage infrastructure managed by third parties.

d. Other expenses

Funding can also be requested to cover:

- Potential travel costs incurred to organize meetings between collaborators from different institutions;
- Costs for open access publication of the project outcomes;
- Participation of the project partners and the SDSC staff to conferences to disseminate the project results.

3. Pre- and Full Proposals documents and procedures

a. Submission details for pre-proposals

The application documents must be submitted before July 13 2020 at 23:59 CEST through the CMT website at https://cmt3.research.microsoft.com/SDSC2020/Submission and must include:

- A pre-proposal (max 5 pages);
- Short CV(s) of all applicants (max 2 pages);
- A list of maximum 10 relevant publications.

The pre-proposal should not only contain the description of the scientific project but should also detail the availability of data and its relevance to the problem considered, provide risk assessments and describe the nature of the collaboration with the center. A detailed template for the pre-proposal document with additional guidelines is available at https://datascience.ch/academic-projects.

b. Submission details for full proposals

If the applicant is invited to submit a full-proposal, the deadline is October 30 2020 at 23:59 CET and the material submitted through the CMT website at https://cmt3.research.microsoft.com/SDSC2020/Submission. The structure of the full application is similar to the pre-proposal and includes:

- An updated and detailed project description (max 15 pages);
- A detailed budget;

- Short CV(s) of all applicants (max 2 pages);
- A list of publications (10 most relevant);
- Supplementary documents, if applicable; e.g., support letters, confirmation of co-operations.

A detailed template for the full-proposal document with additional guidelines is available at https://datascience.ch/academic-projects.

All applicants invited to this step of the procedure can request guidance from SDSC data scientists to help elaborate or refine a proposed data science approach.

4. Contact

For questions regarding the submission of pre-proposals and full proposals, in particular, in the full proposal stage, for questions on the assignment of SDSC staff or on the creation of a detailed budget please contact our Program Manager, Sofiane Sarni, at program.manager@datascience.ch.