Collaborative Data Science Projects

Fifth Call for Projects - 2021

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

Towards that objective, the SDSC publishes an annual funding call for Collaborative Data Science Projects whose aim is to offer the possibility for ETH Domain research groups to benefit from the expertise of the SDSC and to undertake joint research projects with a strong data science component.

So far, about two thirds of the SDSC funding from the ETH Board have been allocated to Collaborative Data Science Projects that are reviewed and selected on a competitive basis. This budget covers (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 applicants' institutions.

The ETH board has recently approved a budget increase dedicated to the improvement of processing and analysis techniques for the growing amounts of data generated from large and complex research infrastructures, sensor networks, and databases at PSI, Empa, WSL, and Eawag.

As a consequence, the current call distinguishes between two possible tracks for collaborative project proposals:

- the general track, dedicated to Data Science for Domain Science, as in previous years
- the new **LSI track**, dedicated to ***Data Science for Large Scale Infrastructures***.

The LSI track is described in Section A.6.

The general track welcomes submissions in all applications domains, including environmental sciences, life sciences including biology and medicine, physical and material sciences, as well as economics and social sciences.

This year project proposals focusing on Personalized Health (PH), that address clinical research questions, are particularly welcome. Such PH projects will receive additional support thanks to a joint initiative between the ETH Domain Strategic Focus Areas SDSC and "Personalized Health and Related Technologies (PHRT)" in order to create synergies in-between the focus areas - which in turn enables us to fund more projects. See Section A.7.

A. Presentation of SDSC Collaborative Data Science Projects

1. Motivation, scope and objectives

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 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 concrete problems requires setting up sizable collaborative projects, which 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 perform analyses to support their research. At the same time, SDSC collaborative schemes aim to support the application of techniques developed in research labs working on data science methods to real world scenarios.

The scope of the 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 and other 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-specific prior knowledge. In a number of applications they need to propagate and predict uncertainties accurately, often in complex structured models, and to be robust to non-stationarities. 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 expertise in a wide variety of data science modeling and computational techniques available to partnering ETH Domain teams. These techniques include statistics, statistical signal processing, machine learning, deep learning, and techniques from computer vision, natural language processing and optimization. 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 methods tailored to the application considered, 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 collaborations 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 also be actively involved in the project, to help guide the development of the methods, and to participate directly in data exploration and analyses.

In most of our projects, the SDSC data scientists working on the project meet on a weekly or bi-weekly basis with collaborators from the partnering teams. These meetings can be at the SDSC, in the partnering lab, or via teleconference. Given the important role of communication and informal exchanges 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 sciences;
- scientific code, implementing methods or analyses, and data libraries, which will be made publicly available whenever this is possible in the context of the project;
- curated and openly accessible datasets (when allowed); and, possibly,
- papers in data science journals or conferences, if the project required specific advances in data science methods.

5. Open science, FAIR principles and Renku

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/.

6. Track on *Data Science for Large Scale Infrastructures* (LSI track)

This track is dedicated to Collaborative Data Science projects that are relevant for the design, operation and exploitation of large and complex research infrastructures, technology platforms, sensor networks, and databases, that are *operated by PSI, Empa, WSL, or Eawag*. These projects can also propose collaborations aimed at improving the processing, the management and the scientific analysis of the data generated in or handled by these infrastructures. More specifically, the eligible infrastructures are:

- PSI: SwissFEL, SLS, SINQ, SmS, CHRISP incl. HIPA, ESI and PANDA
- Empa: NEST, Move and eHub
- WSL: National Forest Inventory (LFI), Long-term Forest Ecosystem Research (LWF) and Avalanche Forecast and Warning
- Eawag: L'EXPLORE and ARES

Applicants from all institutes in the ETH Domain can apply, as long as the infrastructure is eligible. The type of expertise provided by the SDSC is the same as for the general track (see Section A.2).

Proposals whose impact addresses some of the technical challenges and scientific opportunities associated with the growth in data volumes and data acquisition rates are particularly welcome. Note that the data science techniques themselves can but do not need to focus on large-scale machine learning or statistical modeling.

7. Joint initiative on Personalized Health & Medicine with SFA PHRT

Since 2017, Data Science (SDSC) and Personalized Health and Related Technologies (PHRT) are two Strategic Focus Areas (SFA) of the ETH Domain. New technologies developed and utilized in the ETH Domain create ever increasing amounts of new data, such as from the digitization of clinical cohorts at the molecular level or from high throughput imaging of clinical biospecimens. Creating insights based on selected clinical questions from specific data types, or at the interface of different data types, represents a formidable research challenge, but foremost a unique opportunity to gain new (biomedical/clinical) insights, which could support clinical decision-making.

In order to leverage these opportunities the SDSC and PHRT are teaming up to provide in this call more opportunities to leverage Data Science for Personalized Health projects. The number of projects in this area that will be funded in this call will be increased with respect to previous years. SDSC and PHRT will co-fund these additional projects and evaluate all project proposals on PM together.

8. Application procedure overview

The application procedure for Collaborative Data Science projects with the SDSC is divided into two phases: the submission of an initial pre-proposal and, upon acceptance, the submission of a full proposal. The pre-proposal provides a general description of the planned domain science project and of the scientific questions that should be addressed using data science techniques.

At the beginning of September, the SDSC data scientists will reach out to the applicants whose pre-proposals will have been accepted, to offer to meet and discuss their projects. In particular, the objective will be to help define data science questions, modeling approaches and potentially-relevant data science techniques. This interaction is by no means binding and it is at the discretion of the applicants.

The full proposal should provide an in-depth description of the objectives and data science problem. At this stage, in addition to the data science questions, the applicants are encouraged to outline the data science approaches that will be considered. In both documents, the central role of data science in supporting domain science questions should appear clearly, as well as the collaboration scheme with the SDSC.

9. Timeline

Call for collaborative projects	Wednesday June 9th 2021
Pre-proposal **submission deadline**	Monday July 5th 2021
Invitation to submit full proposals	Monday August 9th 2021
Full proposal **submission deadline**	Monday October 25th 2021
Final decisions on full proposals	Monday February 7th 2022
Start of collaborative projects	March-June 2022

B. Eligibility and funded resources

In the remainder of this document, for clarity, the phrase *partnering team(s) (PT),* or *partner(s)* for short, refers to collaborating groups other than the SDSC.

1. General conditions

Eligibility

- 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 activity of their team and manage their resources. These are typically professors and senior scientists.
- Partners can be academic and industrial teams, without restrictions.

Duration

- The project duration is limited to 24 months.

Funding

- Only teams from the ETH Domain can be funded in this call.
- The PI cannot be funded in this call.
- Detailed funding conditions are presented in Section B.3 (see Funded resources).

Submission

- 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 a continuation of already-funded projects from a previous call can be submitted. They will be considered as new proposals and will be subject to the same review criteria as new proposals.
- A proposal can only be submitted to one of the two tracks.
- Resubmissions are welcome.

Follow up

- The PIs of funded projects will be requested to produce written scientific and financial reports every year 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 enforce compliance with standards, good practices, and regulations.

2. Specific conditions for the track on Data Science for Large Scale Infrastructure

- The proposal must relate to eligible large and complex research infrastructures as specified in Section A.6,
- Any researcher employed by an institution of the ETH Domain can apply to this track.

3. Funded resources

SDSC Technical Staff

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

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 (who usually have an interest in data science) for the duration of the project. It is also possible to request funding for a part-time research or computer engineer (e.g. at 10-25%).

The total amount of funding requested for staff hired in the different partnering teams cannot exceed CHF 120k total if there is a single team applying, and cannot exceed CHF 240k total if there are two or more Partnering Teams (and therefore affiliated with different administrative units) from the ETH Domain. There is no restriction on the budget allocated per team within this envelope. The inclusion of a PT in the project should be well justified in terms of complementarity of expertise, skills, and knowledge; the partnering teams are also expected to contribute in kind to the project.

Compute and Storage Resources

The SDSC provides access to compute and storage resources that can be used for the project. To this end, the SDSC partners with national providers (CSCS and SWITCH). Funded projects can have access to compute resources in the form of virtual servers that are tailored to the needs of the project (including GPUs) and/or high performance computing nodes (HPC). Dedicated data storage and user support are also provided. The cost of the compute and storage resources needed by the project is covered in the budget of the project.

Budgeting for these resources will be requested in the annex of the full proposal. You will be provided with templates to choose your configurations and special needs can be discussed. Note that the acquisition of new hardware (e.g. personal computers, sensors, infrastructure, etc.) is not covered, nor is the computing and storage infrastructure managed by third parties.

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 (for those costs not already covered directly by institutions that the teams are affiliated with); and
- Participation of the project partners and of the SDSC staff in conferences to disseminate the project results.

C. Application guidelines

1. Proposal content and structure

The pre-proposal and full proposal should provide:

- A description of the scientific project and of its impact;
- A description of the data that will be acquired/used in the project. The characteristics of the data that are needed to understand and evaluate the feasibility of the project from a data science perspective should be detailed. A data form is supplied to this end;
- A specification of the research questions that should be addressed using data science techniques. To the extent possible, this specification should be formulated in data science terms. Is the objective to learn predictive models? To use methods to identify mechanisms? To answer specific questions using dedicated exploratory data analysis? What are the representations/characteristics of the problem/data that should be used? What are the key elements of domain knowledge that should be taken into account in the design of the models and algorithms?

2. General evaluation criteria

Each proposal is evaluated based on the feasibility of the proposed research project, its relevance, the suitability and nature of the collaboration, and its potential impact. In particular, the following criteria are considered:

- *Clarity of presentation and description of the research questions:* The main challenges and objectives are described with sufficient detail and in terms accessible to non domain experts.
- *Impact and scientific quality:* The proposed research is timely and will potentially have a strong impact. The problem is likely to advance the state of the art in the field.
- *Feasibility, in particular from a data science perspective:* The datasets considered are available and adequate for solving the research question. There is a clear formulation of the problems to solve in data science terms. The stated goals are achievable with the proposed data science approach.
- *Quality of the collaboration:* The roles of all collaborators, including the SDSC, are clearly described and the complementarity of expertise of the different partners is well-justified for the success of the proposed research. This role should not be limited to the provision of infrastructure (e.g. laboratory, compute power, data) but should include scientific and technical contributions needed for the success of the project.
- Promotion of *open and reproducible science* and adherence to FAIR principles.

3. Pre-proposal

Formatting instructions

Pre-proposals should be written in English and cannot exceed 5 pages (A4 size, single spaced) excluding references, and with a minimum font size of 10pt, single line separation, and margins of at least 2.5cm. We recommend using the template provided.

Preparation guidelines

The pre-proposal should give a high-level and accessible overview of the scientific project, from both the domain and the data science perspectives. Along with the pre-proposal, a *data form* should be filled to provide a clear description of the available datasets. If relevant, the applicants have the possibility of submitting on CMT a data sample and/or a metadata file that illustrates some characteristics or structure of the data. The pre-proposals should elaborate as well on the organization, specific roles, and interactions between the different partners and with the SDSC. It should clearly state and describe the precise data science questions that should be the central focus of the collaboration with the SDSC. It should however provide sufficient detail to allow for the assessment of its feasibility and relevance to the current call for collaborative projects. It should highlight what domain science problems will be addressed by data science techniques, without the need to specify precise models and data analyses. It should provide a description of the nature of the collaboration. The main challenges and objectives should be outlined, with specified roles, interactions, and key personnel assigned to each of the tasks. At this stage, however, no detailed work package description or budget allocation is required. A detailed template for the pre-proposal document with additional guidelines and the data-related questionnaire are available at https://datascience.ch/academic-projects.

Evaluation criteria

The main evaluation criteria for pre-proposals are (a) feasibility, as far as it can be assessed from the problem statement in data science terms and the data that the project is based on, (b) the relevance of the expertise expected from the SDSC, and (c) the quality of the collaboration. The pre-proposals are not evaluated in terms of their scientific quality from the domain perspective. A detailed description of potential data science techniques and modeling approaches that address the data science questions is not needed.

The evaluations at the pre-proposal stage are carried out internally by the SDSC. Only those applications that receive a positive evaluation are invited to submit a full proposal during the second phase.

Pre-proposal submission

The application documents must be submitted before July 5th 2021 at 23:59 CEST on the CMT portal (Microsoft's *Conference Management Toolkit*). The link to the CMT submission page will be accessible from https://datascience.ch/academic-projects/ via the link *Pre-proposal Submission* and must include:

- A pre-proposal (max 5 pages);
- Short CV(s) of all applicants (max 2 pages)
- The data form (max 5 pages)
- list of partners to enter directly by filling the form on CMT

Optionally, it is possible to submit on CMT a data sample and/or a metadata file that illustrates some characteristics or structure of the data.

Upon acceptance of the pre-proposal, selected applicants will be invited to suggest the names of three independent reviewers for the evaluation of their full proposal who have both domain science expertise, and, to the extent possible, data science expertise as it applies to their domain.

4. Full proposal

Formatting instructions

Full proposals should be written in English and cannot exceed 15 pages (A4 size, single spaced) excluding references, and with a minimum font size of 10pt, single line separation, and margins of at least 2.5cm. We recommend using the template provided.

Preparation guidelines

Full proposals should contain a detailed description of the proposed research project. It should describe precisely the data science questions, the specific objectives and milestones, possible data science approaches as well as the expected outcomes. An updated version of the *data form* should be submitted along with the full-proposal. The full proposal should elaborate on the organization, specific roles, and interactions between the different partners and with the SDSC. It should also describe the different work packages, with the corresponding leading teams (including the SDSC) and deliverables, along with potential feasibility risks and mitigation strategies. The full proposal should include a detailed budget for the project. A template for the full proposal document with additional guidelines and the budget form will be available at https://datascience.ch/academic-projects.

Evaluation criteria

Full proposals are evaluated based on all criteria listed in Section C.1. both by internal and external reviewers. The project should be presented sufficiently precisely from the domain perspective for a domain expert to be able to assess its scientific quality and impact, while remaining accessible to the non-expert. It is expected that the data science questions will also be specified in more detail than in the pre-proposal and that proposed data science approaches will be described sufficiently precisely to assess their feasibility. The models and techniques should be described whenever possible, but are not systematically needed, as long as a clear roadmap is laid out for the data science approach. The project should nonetheless be positioned with respect to the existing literature at the interface between data science and domain science.

Full proposal submission

If the applicant is invited to submit a full proposal, the deadline is October 25th 2021 at 23:59 CET and the material should be submitted on the CMT portal. The link to the CMT submission page will be accessible from https://datascience.ch/academic-projects/ via the link *Full proposal Submission*.

The structure of the full application is similar to that of the pre-proposal and includes:

- An updated and detailed project description (max 15 pages, excluding references);
- A detailed budget;
- An updated data form (max 5 pages);
- Short CV(s) of all applicants (max 2 pages);
- Supplementary documents, if applicable; e.g., support letters, confirmation of cooperations, data usage agreements, etc.

Optionally, it is possible to submit on CMT a data sample and/or a metadata file that illustrates some characteristics or structure of the data.

Contact

For questions about the submission of pre-proposals and full proposals, please contact our Program Manager, Sofiane Sarni, at program.manager@datascience.ch.