SciLifeLab

Knut and Alice Wallenberg Foundation

The future of life science is data-driven

- Strategy for The SciLifeLab & Wallenberg National Program for Data-Driven Life Science (DDLS)

CHALMERS UNIVERSITY OF TECHNOLOGY · UNIVERSITY OF GOTHENBURG · Karolinska Institutet · KTH VETENSKAP OCH KONST · LINKÖPING UNIVERSITY · LUND UNIVERSITY · Naturhistoriska riksmuseet · SLU · UPPSALA UNIVERSITET · UMEÅ UNIVERSITET · Stockholm University

Table of contents

Executive summary

The future of life science is data-driven, providing major new opportunities to explore and understand biology, human health and changing ecosystems. The SciLifeLab & Wallenberg National Program for Data-Driven Life Science (DDLS) is set up to make use of these opportunities. DDLS is a 12-year research program funded by the Knut and Alice Wallenberg Foundation (KAW) with SEK 3.1 billion (about 300 MEUR). SciLifeLab (Science for Life Laboratory), as a national infrastructure for life science, coordinates this program in close collaboration with ten universities (Chalmers, GU, KI, KTH, LiU, LU, SLU, SU, UmU and UU) and the Swedish Museum of Natural History. Over 12 years, the DDLS program will recruit 39 high-profile young academic leaders, train over 500 PhD students and postdocs, and engage the national research community to apply data science and data-driven research principles and to develop new tools and approaches for e.g. findable, accessible, interoperable and reusable (FAIR) data handling, artificial intelligence, and dynamic models of life. This document describes the motivation, specific aims, and the priorities of the national program as a whole and specifically for the four DDLS research areas; Cellular and Molecular Biology, Evolution and Biodiversity, Epidemiology and Biology of infection, as well as Precision Medicine and Diagnostics. DDLS provides a unique opportunity for Sweden to realize a globally leading role in the future data-driven life science. Furthermore, data-driven research synergizes with the technology-driven SciLifeLab infrastructure engaged in the production of biological data, with national AI and data science programs and with national compute resources will further strengthen this position. DDLS will be based on a strong collaboration across all participating organizations and other life science stakeholders, such as those involved in healthcare, industry and sustainability. The many aspects of the DDLS program will synergize with each other to make Sweden leading in the data-driven life science field as well as to profoundly impact how life science is practiced in the next decade.

Data-driven research is essential for the future of life science in Sweden and globally

Data is the immediate product of research and it is useful not only to its producer, but could also be valuable to other scientists across the world. Life scientists globally are now developing comprehensive collections of data that could be integrated and better utilized to understand biology deeper and more systematically than ever before. For example, the European Bioinformatics Institute manages hundreds of petabytes of public life science data, and this and other data resources are growing rapidly in terms of content, depth and interconnections. This digital trend in life sciences will only accelerate in the future. With the massive increase in computational capabilities and the development of artificial intelligence (AI) and deep learning techniques, researchers are facing a paradigm shift in terms of how life science can be practiced. Data-driven research is also essential for the future of healthcare, for biodiversity, environment and sustainability research as well as for the biotechnology and life science industry. The ability to harness and process data will become a central competitive advantage for both industry and the entire society.

The Swedish government has outlined a national life science strategy, Life Science 2020-2030, aiming to make Sweden a leading life science nation[1]. Data-driven research will play an important role in addressing local and global health challenges. This will require an active engagement of industry, healthcare and other stakeholders. Besides health, data-driven life science will also contribute to solving major societal challenges, such as some of those defined by the UN Global Sustainable Development Goals[2], for example environmental degradation, climate change, and access to sustainable energy, food, clean air and water.

The following specific needs and challenges form the justification for the DDLS program:

1. Major need for data science competence in the life science community

The life science community does not currently have the sufficient multi- and cross-disciplinary skills and competence that the future development of data-driven life science demands. Therefore, global recruitment of talent and training and education of the next-generation of young scientists are essential. Existing experts in bioinformatics and data science will create a critical mass and local research environments for the next-generation scientists.

Overall, to meet these challenges, there is a need to recruit new group leaders globally as well as educate and train new data scientists.

2. Data science capabilities, such as artificial intelligence (AI) are not fully utilized

Recent advances in AI and machine learning (ML) have been dramatic and such approaches are starting to outweigh other approaches to analyze and interpret biological data. For example, AI systems can help to comprehensively analyze the available literature and relevant datasets, contribute to the design, optimization and conduct of laboratory experiments, and help to interpret and visualize data as well as generate insights, conclusions and research reports. Data-driven research is still in its infancy and there will be exponential, even unexpected, developments in the future. Thus, there is a danger that without a dedicated program, the life science community could fall seriously behind, which in turn translates to healthcare, industry, and society at large.

There is a need to i) develop new data analysis and AI methods for life science and ii) promote their availability and use across the broad life science community.

[1] En nationell strategi för life science, Regeringskansliet, artikel nummer N2019.06, https://government.se/information-material/2020/11/swedens-national-life-sciences-strategy/

[2] https://sdgs.un.org/goals

3. Life science data are not shared according to FAIR principles

Despite recommendations, most data are still not readily available according to FAIR[3] principles, and not annotated and organized properly for machine-readability and data analysis. For science to have impact, promoting visibility and awareness of research data is important.

There is a need for coordinated national efforts to facilitate data sharing in life science across the different research areas.

4. Human health data is scattered and not available for research or data-driven precision medicine

Registry data is often used for research, but deeper longitudinal and real-time health data is often not accessible. There is no legislation on secondary use of healthcare data, and a number of legal, ethical and privacy concerns result in the inability to share between healthcare regions and with academia.

There is a need to collaborate across healthcare regions, academia, industry and government to promote health data sharing, as well as to develop innovative technical solutions to fully preserve patient privacy while enabling better utilization of health data locally, nationally and globally.

5. Gaps in the utilization of data in the society

There is often a gap in expertise as well as a lack of access to life science data and the tools and technologies to process it. This was painfully evident during the COVID-19 pandemic, when different sectors of the society would have needed real-time research data across healthcare regions, authorities, industry, the general public and decision makers.

There is a need to ensure that the importance of data-driven insights exists at all levels of the society and that associated legal, regulatory, ethical and policy issues are considered.

Creating a new data-driven, hypothesis generating scientific approach will boost efforts to examine and understand life. This will lead to a powerful iterative cycle, where data science informs on laboratory experiments that in turn feed increasingly accurate, real-time optimized data to computational models of life. This can promote understanding of the basic life science, the prevention, diagnosis and treatment of disease as well as understanding of biodiversity and ecosystem changes. We are still at the start of a digital era for life sciences that will be enhanced and transformed by the capabilities of AI. Together with the 11 partner organizations this truly national DDLS program will be essential in Sweden to lead this transformation and not just react to it, *figure 1*.

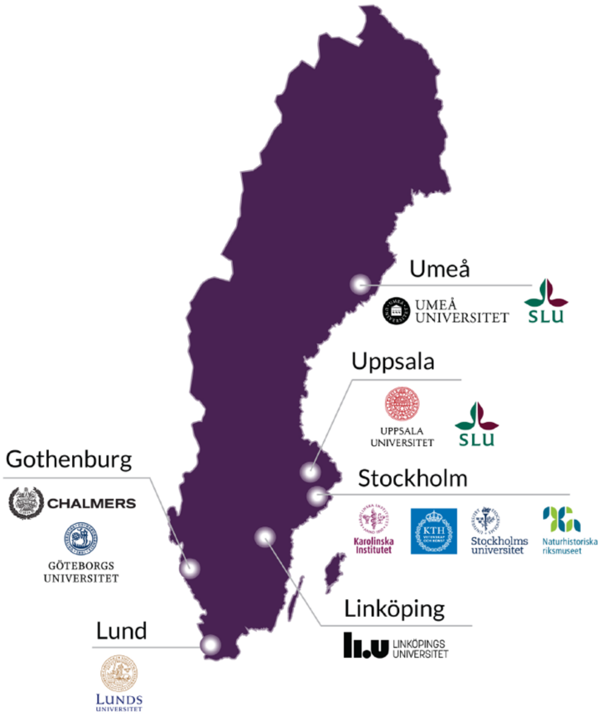


Figure 1 The 11 partner organizations of DDLS: Chalmers University of Technology, University of Gothenburg, Karolinska Institutet, KTH Royal Institute of Technology, Linköping University, Lund University, Stockholm University, Swedish Museum of Natural History, Swedish University of Agricultural Sciences (Umeå and Uppsala), Uppsala University, and Umeå University.

[3] Findable, Accessible, Interoperable and Reusable

Vision and mission

Vision

The future of life science is data-driven

Mission

As a national research and training program across 11 partners, SciLifeLab & Wallenberg National Program for Data-Driven Life Science, DDLS, focuses on how data science and computational approaches enable and energize life science as well as applications in health, environment and industrial research.

Program objectives

The national DDLS mission will be accomplished by recruitment of talent, creation of a national program for training and research excellence, realization of a national data platform with advanced bioinformatic support, and setting up a collaboration hub for engaging academia, healthcare and industry. Therefore, DDLS will have a deep impact for every life scientist, but also for society at large. DDLS aims to change how life science is practised in the future and thus equip the research community with essential future data-driven skills and capabilities.

The program focuses on four strategic areas of data-driven research: Cell and Molecular Biology, Precision medicine and Diagnostics, Evolution and Biodiversity, and Epidemiology and Biology of infections, *figure* 2. All of these are essential for improving the lives of humans, animals and nature, detecting and treating diseases, protecting biodiversity and creating sustainability. To realize the DDLS strategy, the four research areas together with six strategic objectives describe the priorities of this national program, *figure* 3 (*next page*).

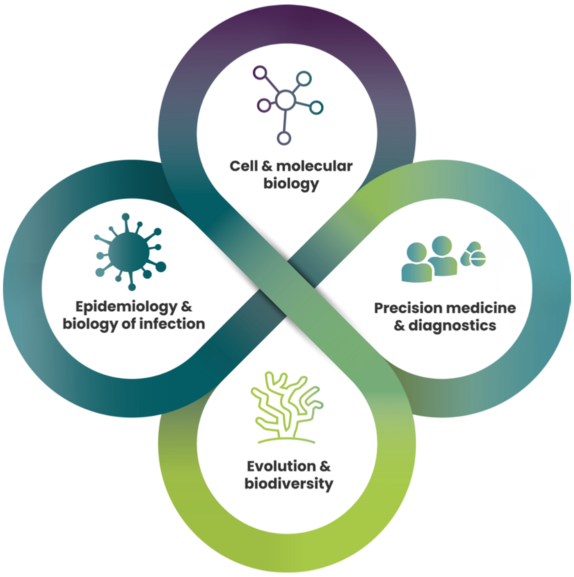
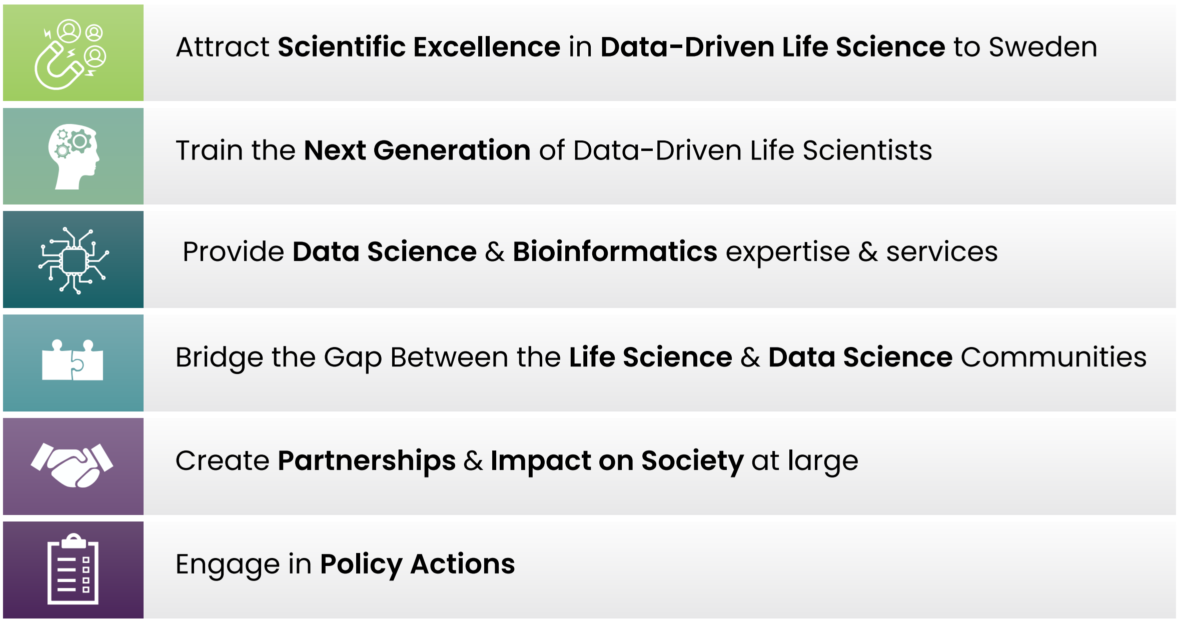

Cell & molecular biology
Epidemiology & biology of infection
Precision medicine & diagnostics
Evolution & biodiversity

Figure 2 Strategic research areas within the DDLS program. *Note: In Swedish the names of the research areas are: cell- och molekylärbiologi, precisionsmedicin och diagnostik, evolution och biodiversitet samt smittspridning och infektionsbiologi.*



Attract **Scientific Excellence** in **Data-Driven Life Science** to Sweden

Train the **Next Generation** of Data-Driven Life Scientists

Provide **Data Science** & **Bioinformatics** expertise & services

Bridge the Gap Between the **Life Science** & **Data Science** Communities

Create **Partnerships** & **Impact on Society** at large

Engage in **Policy Actions**

Figure 3 The six DDLS strategic objectives.

1. Attract scientific excellence in data-driven life science to Sweden

By attracting and recruiting excellence to Sweden, as well as training the next generation of life scientists, DDLS is enabling high-quality and impactful science and equipping Sweden with work force of the future with data-driven skills and capabilities.

The DDLS program is recruiting a total of 39 high-profile young group leaders (tenure-track DDLS fellows) in the four research areas to the participating partner organizations to expand excellence in data-driven life science research. Each fellow is offered a generous starting package with funding covering five years of their salary, and up two PhD student and two postdoc positions, as well as some running costs. After the 5-year fellowship and startup support, the DDLS fellows will have the possibility to be promoted as tenured faculty at their host organization and can then also continue as senior group leaders affiliated with the national DDLS program. We hope that the recruited DDLS fellows will become world-leading experts in their field. All DDLS fellows will be located in progressive, multi-disciplinary local research environments and enjoy powerful links and synergies with all parts of the national DDLS program, as well as with other national fellows programs such as SciLifeLab, WCMM[4] and WASP[5]. For DDLS to succeed in becoming a globally leading program, it is necessary to attract such talented early career scientists, but also to build excellent research environments at each DDLS partner organization. DDLS fellows will be a core to the development of a national DDLS network where long-lasting collaborative interactions between the fellows, their host universities and between different research areas may catalyze powerful synergies. This should have a broad impact on life science, far beyond the individual group leaders. The DDLS fellows will also play an important role in education and training of their students, and in participating in local and national DDLS training events.

Deliverables for objective 1 at the end of 2032:

- DDLS fellows program has led to outstanding international group leader recruitments
- The recruited DDLS fellows have been successful in building strong research groups, launched cutting-edge research projects and have attracted competitive prestigious grants, e.g. ERC.
- Most of the DDLS fellows have been tenured by the host department/unit and have integrated to a progressive local research environment
- The DDLS fellows collaborate with each other, as reflected in joint grants and publications
- The DDLS fellows program has accelerated implementation of data-driven life science in Sweden and increased quality of data-intensive life science publications.

2. Train the next generation of data-driven life scientists

Over the 12 year program period, there will be over 500 PhD students and postdocs who will take part in the DDLS program. Together with the DDLS fellows, the postdocs and PhD students form the core DDLS community.

Besides the PhD students and postdocs that are recruited to the DDLS fellows' groups, we will facilitate global recruitment of top-quality PhD students and postdocs to the DDLS program.

These PhD and postdoc positions will be made available through open competitive calls to qualified mentors working in excellent research environments for data-driven research in the four research areas and across the 11 partner organizations for DDLS. This will help to build a broader national DDLS research community with a critical mass and excellent expertise.

DDLS Research school.
We are launching a national DDLS research school promoting acquisition of high competence and skills in data-driven life science. The research school aims to promote the training, mentoring and collaborations for the 500 PhD students as well as also engage and network the postdoc community. There will be both academic and industry PhDs and postdocs and hence the research school is also promoting academia-industry interactions, facilitating availability of expertise in the industry as well as industrial career options for data scientists. The DDLS research school community will organize meetings and symposia, state-of-the-art training, mentors, as well as collaborations nationally and internationally. DDLS research school will create a strong national network across all students and postdocs.

Courses are being organized together with all DDLS partner organizations as well as other partners and networks, such as the SciLifeLab fellows program, WCMM, WASP, WASP-humanities and society (WASP-HS), the SciLifeLab Bioinformatics Platform and SciLifeLab Data Centre. In addition, SciLifeLab also collaborates with European level training providers, such as EMBO[6], EMBL[7], EBI[8] and ELIXIR[9].

DDLS education and training.
DDLS training and courses will be open to all

[4] Wallenberg Centres for Molecular Medicine
[5] Wallenberg Artificial Intelligence, Autonomous Systems and Software Program

DDLS postdocs and PhD students as well as, when capacity allows, the broader research community. Training is essential to our aim to make the broad life science community better prepared for the opportunities and challenges in the next decade. All senior DDLS funded researchers are expected to take part in the teaching and training in DDLS-branded events and courses. The DDLS training will be built and delivered in a co-creation model with the SciLifeLab training program (SciLifeLab Training Hub) which also maintains tools for "training the trainer" enabling high quality DDLS education and training to further expand.

The training courses and materials will include both foundational data-driven topics and life science specific areas. The training materials will be openly accessible via the SciLifeLab Training Hub to the wider audience, including DDLS researchers, the wide research community, as well as industry and healthcare stakeholders, thus ensuring upskilling the broader community in data-driven knowledge and helping them to implement the data-driven life science knowledge in their research and innovations. The DDLS education and training will also work closely with both national and international collaborators, as explained above for the research school.

Deliverables for objective 2 at the end of 2032:

- Excellent early-stage scientists have been recruited to the program with a high fraction of international students
- A community of more than 500 PhD students and postdocs has been trained.
- A unique national research program and collaborative network across the 11 DDLS partners, collaborators, industry and healthcare is created.
- A broad collection of training materials exists for various level of expertise (novice to expert) covering a broad range of data-driven life science topics.

3. Provide data science and bioinformatics expertise and services

To increase the impact and reproducibility of research data as well as cutting-edge data-driven research we will create an environment where researchers, data-producing platforms, advanced bioinformatics support and compute centers interact.

SciLifeLab data platform.
We will build a national DDLS data platform and e-infrastructure with a range of advanced data analytics support services that are integrated into the research environment and widely available to all life science researchers and data-producing facilities in Sweden. The platform will also facilitate access to top computational resources. The platform will create data resources and services, and organize scientific information, including data, code, methods, and meta-data in a reusable way. The central technical hub is operated by the SciLifeLab Data Centre. This will operate the central resources and coordinating activities at the national

[6] Excellence in life sciences (https://www.embo.org)
[7] European Molecular Biology Laboratory (https://www.embl.org)
[8] EMBL's European Bioinformatics Institute (https://www.ebi.ac.uk)
[9] ELIXIR |A distributed infrastructure for life-science information (elixir-europe.org)

level. Thus, the data platform links together data, researchers and e-infrastructures as well as advanced data analytics and bioinformatics support, all embedded in the local research- and data environments at the host organizations. Four Data Science Nodes (DSNs) are launched, one for each of the DDLS strategic research areas. DSNs provide national data services, data analysis and bioinformatics support and coordinate data-related activities across the country.

Advanced bioinformatics support is available for researchers and provided via the Wallenberg Advanced Bioinformatics Infrastructure (WABI) at the SciLifeLab Bioinformatics Platform, NBIS[10]. By co-locating and integrating data platform staff, software developers, data engineers, data scientists, data stewards and bioinformaticians at the DSNs into strong environments for data-driven life science and data infrastructure, we think that the future capabilities of Swedish data-driven life science are substantially facilitated.

Recruitment of bioinformatics experts.
The DDLS program will recruit senior bioinformatics experts to deliver cutting-edge competences and services to the DDLS community. The staff will be nationally and internationally recognized bioinformatics experts and will be part of the DDLS DSNs and the support teams at the SciLifeLab bioinformatics platform, NBIS. The experts will be tasked to ensure that state-of-the-art bioinformatics know-how is efficiently maintained, applied and spread to the research community, thereby creating high-quality analysis capabilities of complex data.

Making data FAIR.
The DDLS program coordinates efforts across the country to make FAIR data the norm in academia, as well as creating services and resources to support this. The data platform and the DSNs support the needs of the main scientific areas of the DDLS program, and establish research area specific portals, following the example of the national COVID-19 portal with links to international data resources.

Connecting high-performance computational capabilities.
The DDLS program will link up with national powerful computational capabilities to develop new computational methods to facilitate data-driven life science research. The KAW-funded Berzelius GPU HPC as well as the new national super computing capabilities coordinated from LiU, advanced data analysis technologies and AI capabilities are resources that will be connected to the program. The program is also promoting research that contributes to international key projects of future compute services in the domains of life science (e.g. in collaboration with EBI, EMBL Data Science Center, ELIXIR as well as efforts regarding for example the European Health Data Space.)

Deliverables for objective 3 at the end of 2032:

- Established a national data platform, broadly enabling and facilitating FAIR data sharing
- Create domain-specific data resources and portals in the four research areas with the DSNs
- Developed high-end computational and ML/AI technologies to transform life science
- Set up advanced computational services for the whole life science sector
- Profile the DDLS program as an internationally recognised role model for how bioinformatics know-how is efficiently maintained, applied and spread in the research community, by the recruitment of leading bioinformatics support experts.

[10] National Bioinformatics Infrastructure Sweden (https://nbis.se/)

4. Bridge the gap between the life science and data science communities

DDLS is a multi-disciplinary research program where the key aim is to bring life scientists and data scientists to interact and to increase knowledge of data science among life scientists and vice versa. DDLS also aims to stimulate cross-disciplinary collaborations between life scientists and data scientists. Dedicated funding, joint calls and networking events will be carried out to enhance such interactions.

Collaboration with WASP.
Wallenberg AI, Autonomous Systems and Software Program (WASP) is a major research program which provides a platform for academic research and education, fostering interactions with leading Swedish companies. The WASP program focuses on research in AI, autonomous systems, and software as enabling technologies for developing systems acting in collaboration with humans, systems that adapt to their environment through sensors, information, and knowledge, and thus form intelligent systems. Collaborations with WASP enable the life science community to interact and synergize with the large community of world-leading AI scientists and software- and automation experts. Additionally, this will also introduce life science challenges to the WASP community. WASP has powerful research methodologies focusing on e.g. mathematics, machine learning, or autonomous systems and software. DDLS will launch joint calls with WASP such that each party funds scientists in their own community. The calls are open for and must engage both a DDLS associated research group and a WASP research group and include relevant scientific challenge for both sides.

Deliverables for objective 4 at the end of 2032:

- 50 joint research projects between DDLS and WASP communities have been funded, significant high impact publications published and long-term collaborations established.
- The DDLS fellow communities have regular joint meetings with WASP, SciLifeLab, WCMM and other fellow programs.

5. Create partnerships and impact on society at large

Addressing the grand challenges related to human health and the environment requires transdisciplinary efforts between many scientific disciplines, in particular between life and data sciences, and across the sectors of academia, industry, and healthcare. DDLS will boost data-driven life science through partnerships nationally and with the global research community.

Fellows' communities.
Besides the WASP community, DDLS is collaborating with other research programs that are focussing on fellows, such as the SciLifeLab Fellows' program, the Wallenberg Centers for Molecular Medicine (WCMM), and the Wallenberg Centre for Quantum Technology. The DDLS fellows will have the opportunity to take part in the Program for Academic Leaders in Life Science (PALS), which

is a career and scientific collaboration program bringing together three fellows' programs in life science, the SciLifeLab and DDLS fellows programs as well as the WCMM centres. PALS program also funds small research grants to start collaborations across PIs in these communities. This will provide DDLS group leaders excellent opportunities to collaborate with experimental and clinical researchers.

Industry PhD and postdocs.

At least 90 industry PhDs and postdocs will be recruited to facilitate collaborations between academia and industry. They will be enrolled in the DDLS Research school, education and training activities. Industry is an important beneficiary of the DDLS training programs and DDLS aims to facilitate knowledge transfer and career opportunities between academia, industry and the public sector.

DDLS innovations.

DDLS will support innovation processes through collaborations with industry and healthcare, largely via making use of existing innovation support systems. One example is the collaboration with the KAW innovation program, the Wallenberg Launchpad (WALP) program. WALP can provide proof of concept funding to develop innovative ideas from KAW funded research. Through the WALP program, DDLS researchers will gain a bridge towards venture funding to develop their projects and products towards the markets.

Partnering with healthcare.

DDLS training will help to create a community of scientists that may have careers as next-generation data experts in the healthcare. In the research areas of Precision Medicine and Diagnostics as well as for Epidemiology and biology of infection, the program seeks to integrate molecular and clinical data in a research setting. DDLS is also working together with the healthcare regions, biobanks, Genomic Medicine Sweden, and the WCMM network to promote data science in health research. In order to develop closer links with healthcare, it would be advisable that the academic DDLS partner universities create opportunities for DDLS fellows to collaborate with the clinics and to undertake joint positions in the healthcare.

Synergies with the global research community.

DDLS promotes access and availability to international training programs for the Swedish research community. International networking is key to the success of DDLS, and we are building collaborative programs together with leading international institutions in data-driven research, such as through EU programs (e.g. through several funded projects in the European Health Data Space) and Nordic collaborations. We already have collaborations underway with the EMBL, and we will work with partners to create and promote international standards and practices in data handling. In addition, DDLS is engaging with other national communities within research areas such as biodiversity, environment, agriculture and forestry.

SciLifeLab infrastructure units and their user community of about 2000 scientists is an important interaction opportunity for the DDLS program and the DDLS community. The combination of technology-driven and data-driven science is a unique opportunity for SciLifeLab. DDLS will work with the infrastructure units and SciLifeLab Data Centre to promote the FAIR-data principles as an integral part of the national infrastructure operations.

Deliverables for objective 5 at the end of 2032:

- The DDLS group leaders and scientists have extensively networked with other national research communities, such as SciLifeLab fellows and PIs, WASP and WCMM and others as evidenced by joint publications, grants and high-level participation in annual PALS meetings.
- Extensive collaborations have taken place with industry, healthcare, and other national stakeholders.
- Overall, an active DDLS community has been formed involving both academic researchers, industry, healthcare and other relevant actors from both public and private sectors.
- There are many examples where DDLS research has been translated and developed into commercial products and services, or disseminated in other ways thus demonstrating significant impact and benefit for the society.
- Collaborative interactions of DDLS with leading international institutions and networks have resulted in scientist exchanges, joint grants, joint symposia and collaborations evidenced by joint publications.

6. Engage in policy actions

Effective and responsible data-driven research may require policy actions and attention to ethical, legal and social aspects of research. DDLS will engage in these matters as issues arise, often in collaboration with other stakeholders. These issues may involve access to healthcare data, secondary use of healthcare data, privacy issues, the challenges and ethical issues in the use of AI etc. On many occasions, practical implementation of data-driven research is dependent on solving such society issues.

Collaboration with WASP-HS.

The Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) is a research program in which researchers tackle the challenges and impacts of the upcoming technology shifts. The program will contribute to developing theories and practices concerning human and societal aspects of AI and autonomous systems, emphasizing on the ethical, economic, labor market, social, cultural, and legal aspects of the technological transition. We will work to bridge the DDLS and WASP-HS communities and foster collaborative projects with aim to investigate human and social challenges of data-driven strategies developed within the life sciences. Focus will be on the four DDLS research areas and areas related to AI and autonomous systems, or data as well as the ethical, legal, social, economic, cultural or policy issues that arise from their application.

Ethical, legal and social aspects of data-driven research.

Progress in many areas of life science is highly dependent on ethical, legal, and social implications (ELSI) as well as various regulations and practical limitations. These include data security, privacy, ownership, sharing, fragmentation of, and access to, healthcare data, all of which are already actively discussed at the national level and internationally. In addition, implications (ELSI) as well as various regulations and practical limitations. These include data security, privacy, ownership, sharing, fragmentation of, and access to, healthcare data, all of which are already actively discussed at the national level and internationally. In addition, questions related to policies in biodiversity and sustainability of environmental research also need to be addressed, e.g. benefit sharing. The DDLS program is working with the broader community of stakeholders to connect leading experts on ELSI and related matters to the program. A policy action group will be formed to identify questions and roadblocks to DDLS research and to the progression to a digital, data-driven future in life science research.

Deliverables for objective 6 at the end of 2032:

- DDLS actions and those of other stakeholders have helped to move questions concerning health data access and sharing forward in Sweden.
- DDLS has influenced national debates on biodiversity and environment.
- Challenges in policy and ELSI issues in DDLS research have been catalogued and individual solutions identified.

National research programs: four research areas and 11 partner organizations

As DDLS is anchored at 11 partner organizations across the country, the formation of active national collaborative communities will be key for the success of the program and its ability to reach to an international level of scientific excellence. It is also a unique opportunity to achieve integration of research activities in data-driven life science across universities. By forming synergistic national networks we expect to see collaborations within each DDLS research area, including data suppport and computational capabilities, but also cross-disciplinary collaborations across the four research areas. In this way ground-breaking research can develop bottom-up from within the data-driven life science community. The 39 DDLS fellows will form the base for building a core community and will participate in research collaborations and training, *figure* 4.

The four DDLS research areas will all work towards the six long-term objectives of the DDLS program as described in the section *Program objectives*. Each research area has its own profile and niche opportunities but also challenges that will need to be addressed as policy and/or ELSI matters. These four research areas are further explained below:

Cell and molecular biology

The subject area concerns research that fundamentally transforms our knowledge about how cells function by analyzing data from their molecular components in time and space, from single molecules to native tissue environments. This research area aims to lead the development or application of novel data- driven methods relying on machine learning, artificial intelligence, or other computational techniques to analyze, integrate and make sense of cell and molecular data.

Precision medicine and diagnostics

The subject area concerns research that will make use of computational tools to integrate molecular and clinical data for precision medicine and diagnostic development. The focus is on data integration, analysis, visualization, and data interpretation for patient stratification, discovery of biomarkers for disease risks, diagnosis, drug response and monitoring of health. The precision medicine research is expected to contribute with strong capabilities in machine learning, AI and other computational tools to make use of existing strong assets in Sweden, such as molecular data (e.g. omics), imaging, electronic healthcare records, longitudinal patient and population registries, biobanks and digital monitoring data.

Evolution and biodiversity

The subject area concerns research that takes advantage of the massive data streams offered by techniques such as high-throughput sequencing of genomes and biomes, continuous recording of video and audio in the wild, high-throughput imaging of biological specimens, and large-scale remote monitoring of organisms or habitats. This research area aims to lead the development or application of novel methods relying on machine learning, artificial intelligence, or other computational techniques to analyze these data and take advantage of such methods in addressing major scientific questions in evolution and biodiversity.

Epidemiology and biology of infection

The subject area concerns research that will use big experimental, clinical, or pathogen surveillance data in innovative ways to transform our understanding of pathogens, their interactions with hosts and the environment, and how they are transmitted through populations. The priority area covers computational analysis or predictive modelling of pathogen-host systems for which multidimensional, genome-scale experimental data are now available and extends to using population-scale genetic, clinical, or public health data from pathogen surveillance efforts and biobanks.

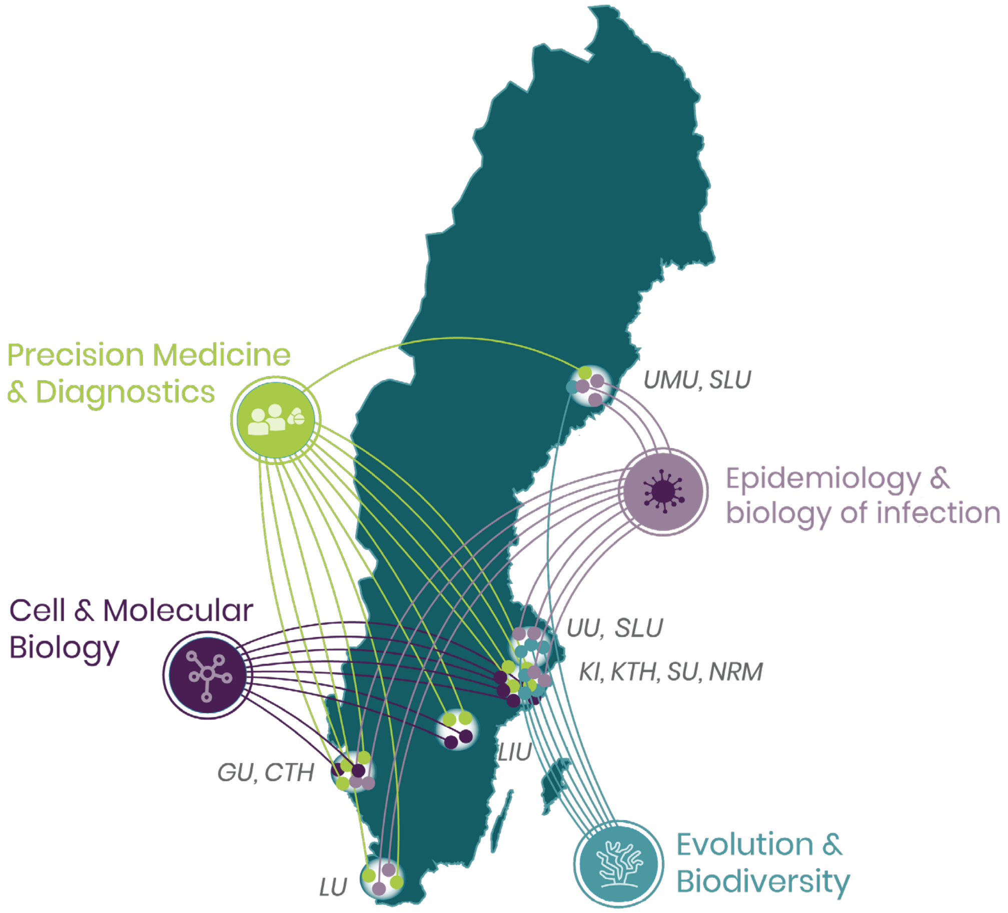


Figure 4 Distribution of DDLS Fellows positions according to the donation letter from KAW.

How do we define data-driven life science? Funding principles and expectations

During the launch phase of the DDLS program, it became clear that we need to define data-driven research in such a way that scientists and stakeholders interpret the DDLS goals in the same way.

Data-driven life science is a field of research that focuses on using data, computational methods and artificial intelligence to study biological systems and processes. This approach includes assembling, sharing, integration and advanced analysis of large amounts of data from diverse sources, including experiments, observations, and simulations, in order to gain a better understanding of how living organisms function.

In addition, we will consider the following aspects when DDLS funding decisions are taken.

- DDLS projects are engaged in the application of data science to solve concrete questions in basic and applied life sciences, in the four research areas. This includes development of new data science methods but also the intelligent adaptation of existing methods to studies of biology. Basic data science, without specific biological questions, is not a core activity for DDLS. However, through the DDLS-WASP collaborative program, DDLS promotes interdisciplinary interaction of life science researchers with data scientist and AI experts.
- DDLS will not support projects that are primarily laboratory-based where the generation of large-scale exploratory data sets is the main aim (e.g. via omics profiling or imaging). However, targeted laboratory work to validate hypotheses gained from data-driven research is an important part of DDLS.
- DDLS will not only deal with publicly available data. Projects engaged with new and unique datasets are highly valued, but ideally the experiments and data generation should be accomplished using other funding.
- Datasets applied in DDLS research are expected to be made openly accessible according to the FAIR principles.
- Algorithms and code that are developed during DDLS research should be properly deposited and made available. Publication in open access journals is expected to the extent possible. These practices do not preclude seeking patent or IP protection for innovations. Also, separate agreements will be created for industrial collaborations.
- DDLS is an academic program, hence contributions to scientific excellence are expected in all projects, including industrial and healthcare collaborations.
- DDLS values education and training. Hence, funded group leaders and project participants are expected to take part in the DDLS training and educational events and activities.
- DDLS recruitment and funding decisions should favor young scientists, international recruitment of talent and take into account diversity, equity, and inclusion aspects.
- DDLS expects funded research to be carried out according to good scientific practice, under appropriate ethical approvals for human or animal studies and protecting sensitive private data.
- In addition, project specific conditions for funding will be applied.

SciLifeLab as national host for the DDLS program

SciLifeLab (Science for Life Laboratory), as a national infrastructure for life science, coordinates the DDLS program in close collaboration with the ten universities and the Swedish Museum of Natural History. DDLS will gain substantial synergies from SciLifeLab as a technology-driven national infrastructure provider, and as a major generator of life science data nationally and internationally. We will build on the national SciLifeLab organization in the data support (e.g. the SciLifeLab Data Centre and the Bioinformatics platform NBIS) as well as make use of SciLifeLab Operations Office in the coordination and administration of the DDLS program, such as communication, external relations, training, meetings, events, financing and reporting.

With the launch of the DDLS, SciLifeLab will now have the opportunity to combine technology- and data-driven approaches at the national level and collaborate with all the 11 partners and the four research areas. The establishment of SciLifeLab sites around the country, and the launch of SciLifeLab capabilities in Pandemic Laboratory Preparedness, Precision Medicine and Planetary Biology synchronize well with the DDLS research areas and we see strong added value to the DDLS program and life science in general in Sweden.

Concluding remarks

The DDLS research program is an extraordinary opportunity for scientists in Sweden to come together and build a new generation of data-driven research capabilities as well as recruit and train a new generation of multi-disciplinary data-driven scientists that academia, industry and society need. DDLS will address some of the biggest opportunities that science has to offer based on deep analyses of global and local research data. We will promote open science and FAIR data with concrete actions.

Success will depend on how well we manage to integrate the six program objectives, the four research area topics, the 11 partner organizations and numerous collaborative communities, such as industry, healthcare and society sectors. We can only guess where the exponentially developing computational capabilities, AI and data science will take us in the next decade, but we do hope that with the help of the DDLS program, Sweden and the swedish scientists will have a critical mass of expertise to act in the driver's seat as life science undergoes a new phase.

Acknowledgments

The strategy for the DDLS program was developed by the DDLS steering group with input from KAW, the 11 participating organizations, as well as the SciLifeLab Board, SciLifeLab International Advisory Board, SciLifeLab Management Group, SciLifeLab Operations Office, SciLifeLab Data Centre, NBIS, and in dialogue with other KAW-funded research programs WASP, WASP-HS and WCMM.

We are also grateful for input from organizations participating in the open digital meeting, "The future of life science is data-driven – DDLS update and strategy"[11] on September 14, 2021 when the DDLS strategy was presented and discussed. This is the second version of the strategy for SciLifeLab & Wallenberg National Program for Data-Driven Life Science (DDLS), released 2023.

Creating a joint long-term strategy with all stakeholders will enable the ambitions of the national DDLS program to succeed and create synergistic benefits for the entire life science ecosystem thus create capabilities to tackle the future societal grand challenges.

Finally we acknowledge Knut and Alice Wallenberg foundation for their generous donation enabling the SciLifeLab & Wallenberg Data Driven Life Science Program (Donation letter: KAW 2020.0239).

[11] https://www.youtube.com/watch?v=LkMSgdrNXgo