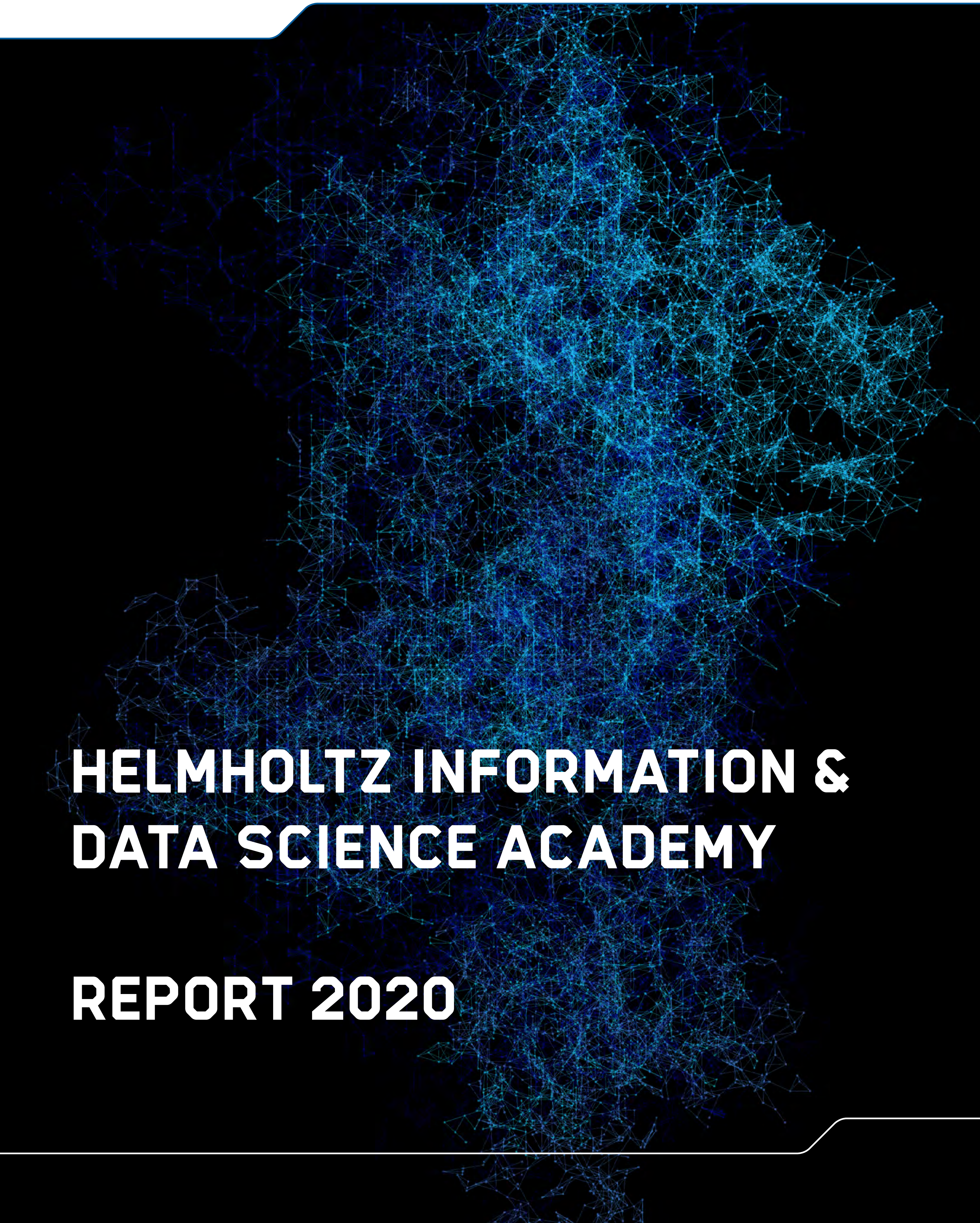
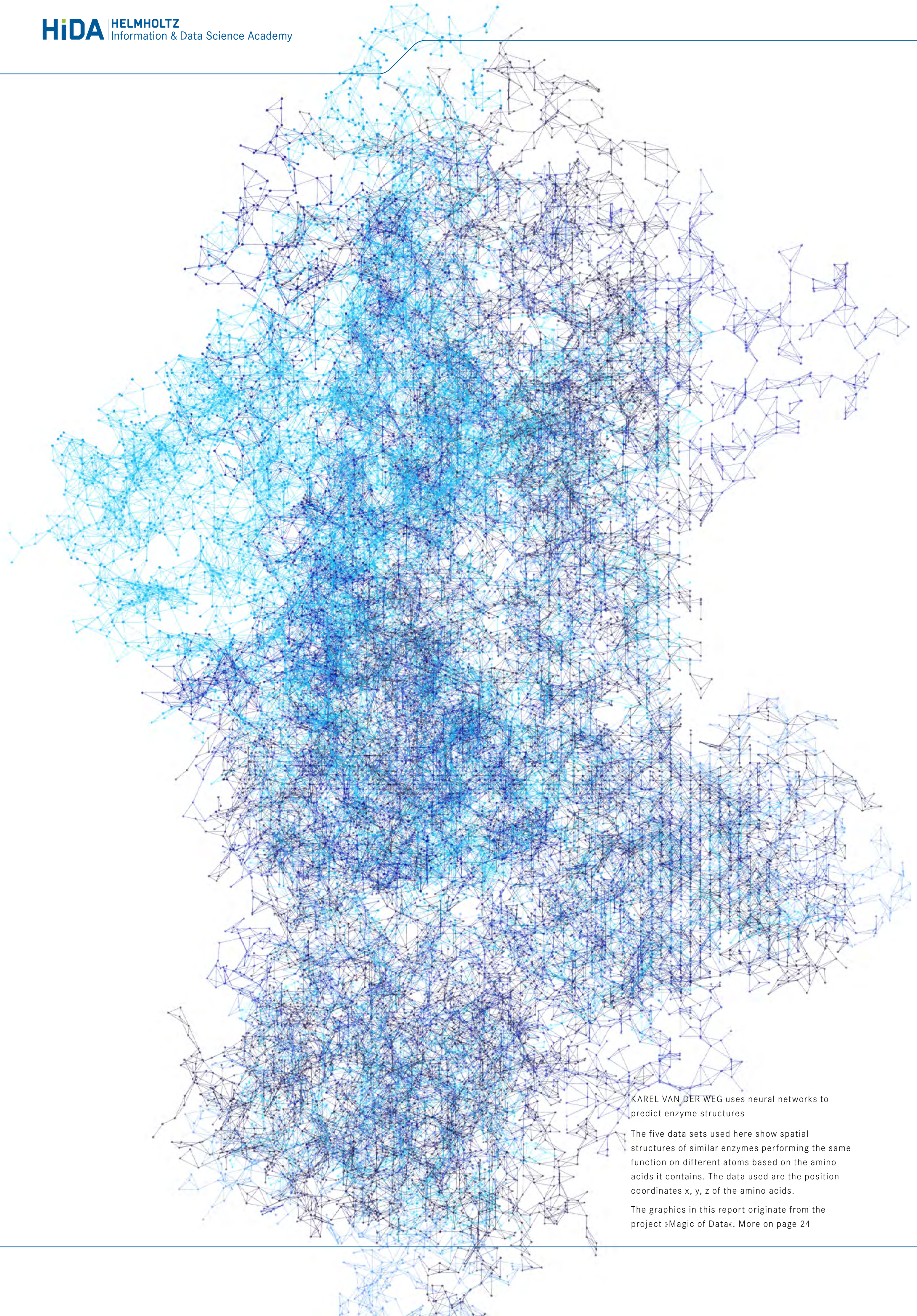
HiDA HELMHOLTZ
Information & Data Science Academy



HELMHOLTZ INFORMATION & DATA SCIENCE ACADEMY

REPORT 2020



KAREL VAN DER WEG uses neural networks to predict enzyme structures

The five data sets used here show spatial structures of similar enzymes performing the same function on different atoms based on the amino acids it contains. The data used are the position coordinates x, y, z of the amino acids.

The graphics in this report originate from the project »Magic of Data«. More on page 24

I. HIDA IN GENERAL

A central component of attracting and promoting excellent young scientists in the field of Information & Data Science is the Helmholtz Information & Data Science Academy (HIDA), founded in 2019. It provides networking activities, further education and training of data science talents who combine knowledge in state-of-the-art information processing with know-how in a scientific research field. HIDA seeks to support the recruitment of the most talented young scientists in the field worldwide for the Helmholtz Centers, thereby substantially contributing to the employer branding to the benefit of all of Helmholtz' research programs. In full extension, HIDA and the Helmholtz Information & Data Science Schools (HIDSS) will train over 280 doctoral researchers until 2025 (incl. associated doctoral researchers as reported by the schools). This makes the program the largest structured postgraduate training program in the field of Information & Data Sciences in Germany.

In 2020, HIDA focused on establishing a rapidly growing network between the 18 Helmholtz Centers, numerous top universities and other partners with expertise in the field of Information & Data Science, thus creating a pathway to successfully recruit and train young talent. HIDA has provided a range of virtual training and networking opportunities in the field of Information & Data Science - with the aim of further intensifying collaboration between researchers, creating spaces for exchange and teaching, and actively supporting the Helmholtz Centers in recruiting new data science talent.

By training top talent in the HIDA network - a large proportion of the researchers trained here will go on to work in industry - the Helmholtz Association is making a substantial contribution to closing the skills gap in the field of Information & Data Science and thus making a further strong contribution to the future viability of Germany as a place for innovation.

II. HELMHOLTZ INFORMATION & DATA SCIENCE SCHOOLS

The HIDA network is the largest postgraduate training program in Information & Data Science in Germany. At the core of this program are the six Helmholtz Information & Data Science Schools, which form a network between 14 Helmholtz Centers and 17 top-tier universities that will train over 280 doctoral researchers (incl. associated doctoral researchers) until 2025. The schools are groundbreaking in the development of new collaborative approaches to evaluating complex, heterogeneous data in the natural sciences with the help of intelligent algorithms - thus enabling modern cutting-edge research.

Despite the COVID-19 restrictions, the six schools under the roof of HIDA continued to expand their activities with great success in the reporting year 2020. Currently, 169 doctoral researchers are being trained at the schools in Hamburg, Kiel, Heidelberg/Karlsruhe, Berlin, Munich and Jülich.

In 2020, **46 first-author and 29 co-author publications** were written and published at all schools combined.

The schools' customized curricula offer a wide range of lectures, seminars, workshops, trainings and summer schools. In 2020, most of these took place virtually:

74 lectures were held at all schools (with the additional MUDS Seminar Series that took place at two-week intervals since August 2020). The lectures are both specific to the subject matter of the schools (e.g. "Beyond Just Fitting Numbers: Interpretable Artificial Intelligence for the Small Data of Materials Science" at HEIBRiDS) and broader in scope (e.g. "Introduction to Visual Data Science" at HIDSS4Health). They also reflect very recent developments as shown by the example of the "Hamburg COVID 19 Lecture Series" held at DASHH in summer 2020.

In total, **27 courses** took place across all schools in 2020. They were aimed at providing doctoral researchers with specific skills, exemplified by courses like "Insights into Python" (MarDATA) or "Good scientific practice" (HDS-LEE). Also to be mentioned are the two Journal Club Series of HIDSS4Health in February and October 2020.

63 networking events of various types were organized by the schools, including PhD seminars. This figure also contains events that are part of the schools' regular program, such as MarDATA's Digital Science Mondays.

RECRUITMENT

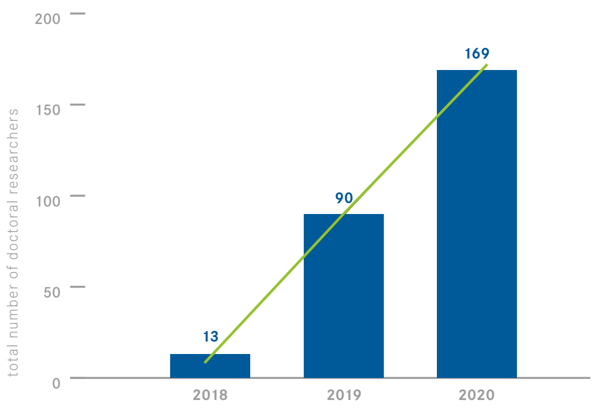


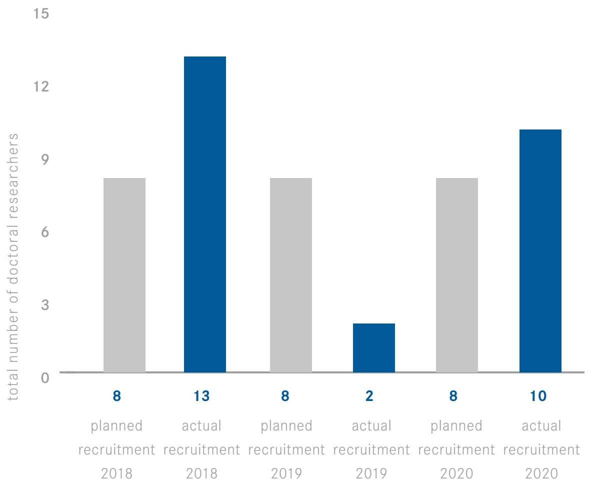
Fig.1: Recruitment of doctoral researchers across the schools (incl. associated doctoral researchers).

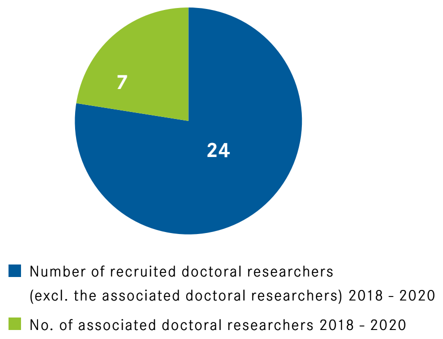
Figure 1 shows the total number of recruited doctoral researchers from 2018 up to 2020. The graph also shows that the number of recruited doctoral researchers increased steadily from 13 doctoral researchers in 2018 to 169 doctoral researchers in 2020 (incl. 46 associated doctoral researchers, who are funded by third parties or other programs). The recruiting activities are according to plan and show the steady growth of the schools.

RECRUITMENT OF DOCTORAL RESEARCHERS PER SCHOOL

1. HEIBRIDS

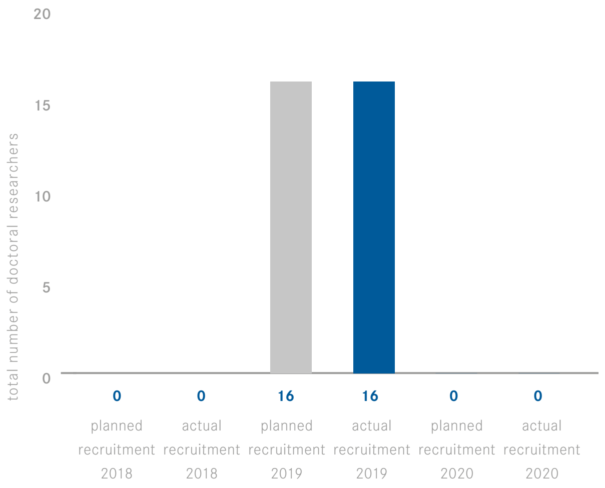
Fig. 2 & 3: Recruitment of doctoral researchers at HEIBRiDS: Figure 2 shows the number of recruited doctoral researchers (excl. the associated doctoral researchers). In total 31 doctoral researchers were recruited within the last three years (incl. associated doctoral researchers). One doctoral researcher, who was recruited in 2018 became an associated doctoral researcher in 2020. In figure 3 he is listed as associated doctoral researcher.

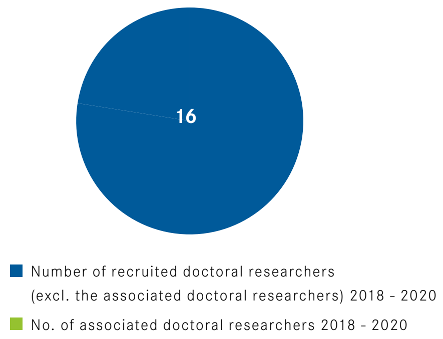




2. MARDATA

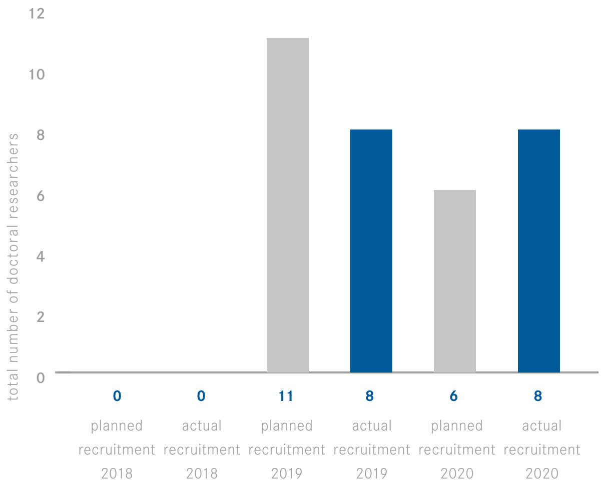
Fig. 4 & 5: Recruitment of doctoral researchers at MarDATA: As planned, 16 doctoral researchers were recruited so far. In 2020, MarDATA didn't plan any recruitment round. This is in line with the concept.

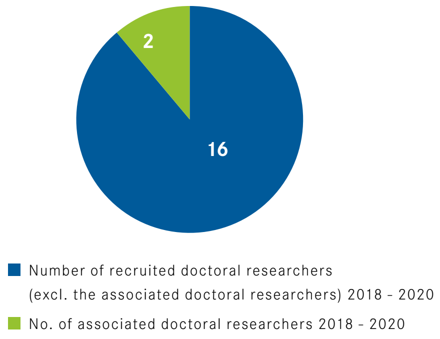




3. DASHH

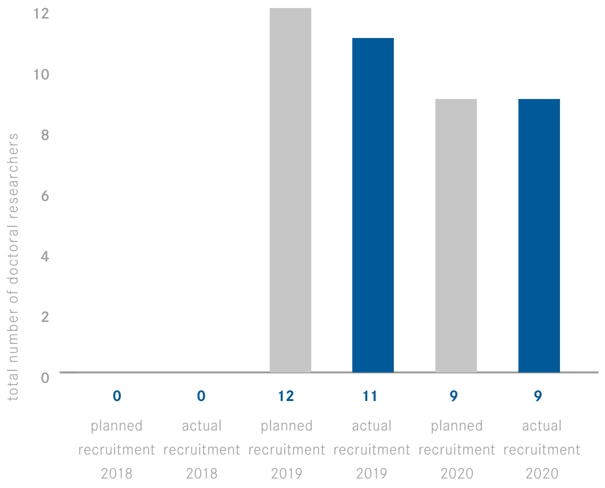
Fig. 6 & 7: Recruitment of doctoral researchers at DASHH: Figure 6 shows the recruited doctoral researchers (excl. associated doctoral researchers). As figure 7 shows 18 doctoral researchers in total were recruited within the last three years (incl. associated doctoral researchers).

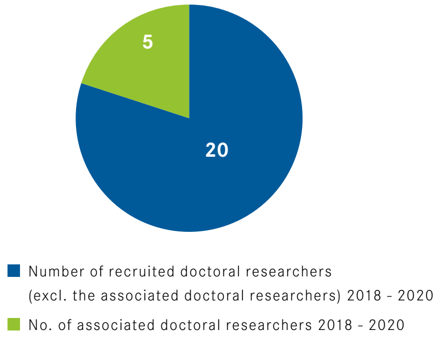




4. HIDSS4HEALTH

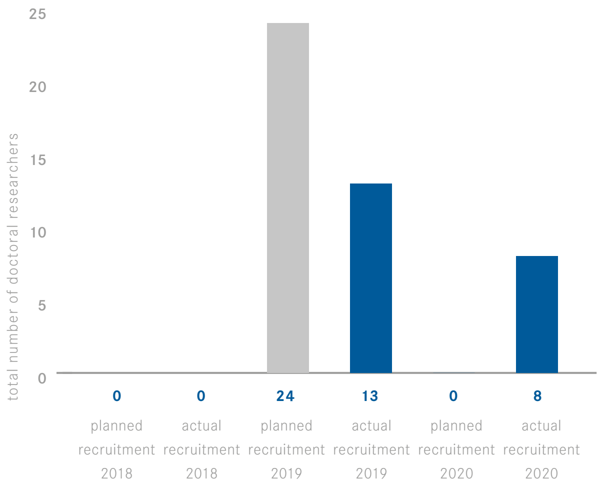
Fig. 8 & 9: Recruitment of doctoral researchers at HIDSS4Health: Figure 8 shows the recruited doctoral researchers (excl. associated doctoral researchers). As figure 9 shows in total 25 doctoral researchers were recruited within the last two years (incl. associated doctoral researchers).

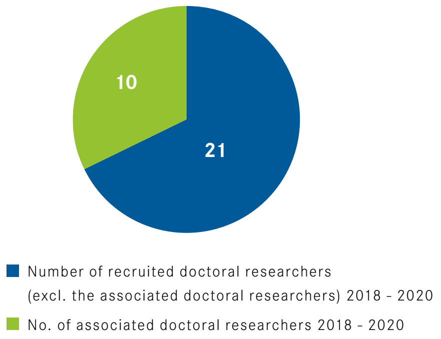




5. HDS-LEE

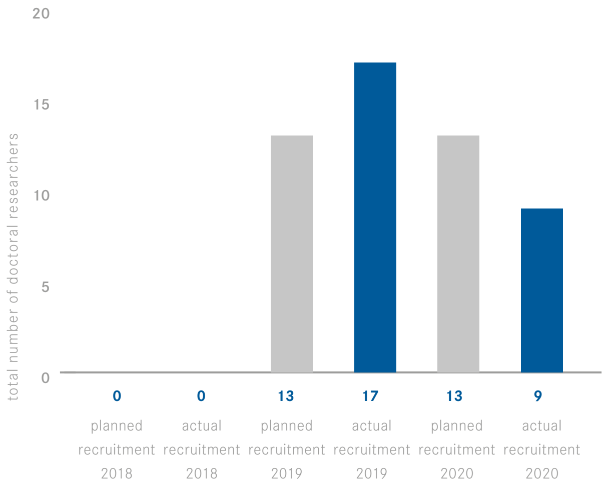
Fig. 10 & 11: Recruitment of doctoral researchers at HDS-LEE: The school originally planned to have one cohort starting together in 2019, but it changed plans and organized two more recruitment rounds. Figure 10 shows the recruited doctoral researchers (excl. associated doctoral researchers). As figure 11 shows in total 31 doctoral researchers were recruited within the last two years.

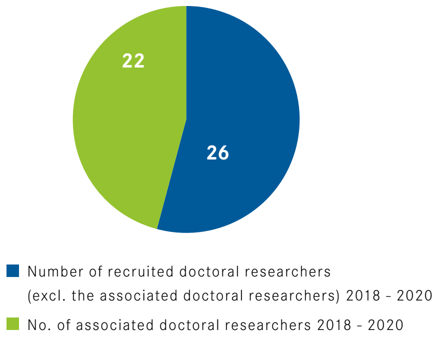




6. MUDS

Fig. 12 & 13: Recruitment of doctoral researchers at MUDS: Figure 12 shows the recruited doctoral researchers (excl. associated doctoral researchers). As figure 13 shows 48 doctoral researchers in total were recruited within the last two years.





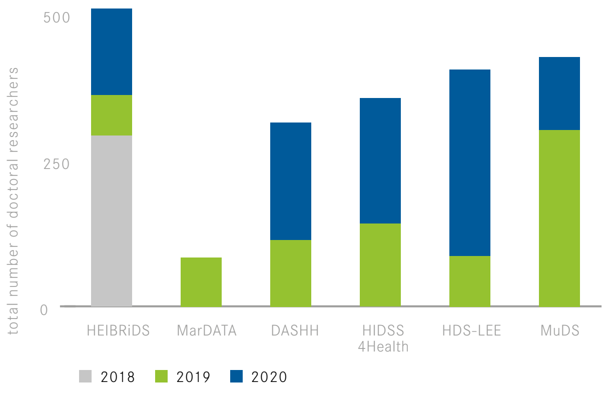


Fig. 14: Number of received applications listed by school.

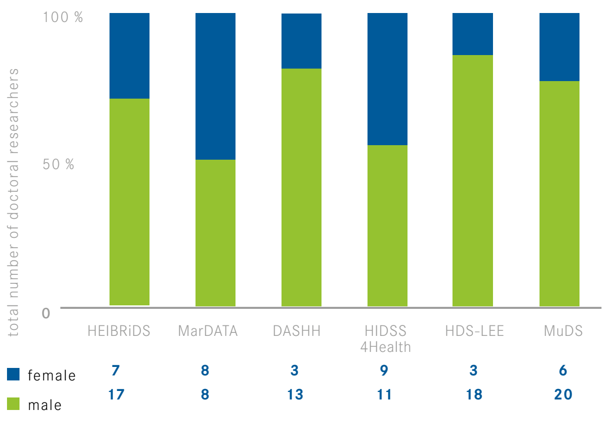


Fig. 16: Gender ratio of the doctoral researchers across all schools (excl. associated doctoral researchers).

The schools made enormous recruiting efforts, which were considerably supported by HIDA. Figure 14 shows the number of applications the schools received from 2018 to 2020. In total 2173 talent applications from around the world were submitted to the Helmholtz Information & Data Science Schools.

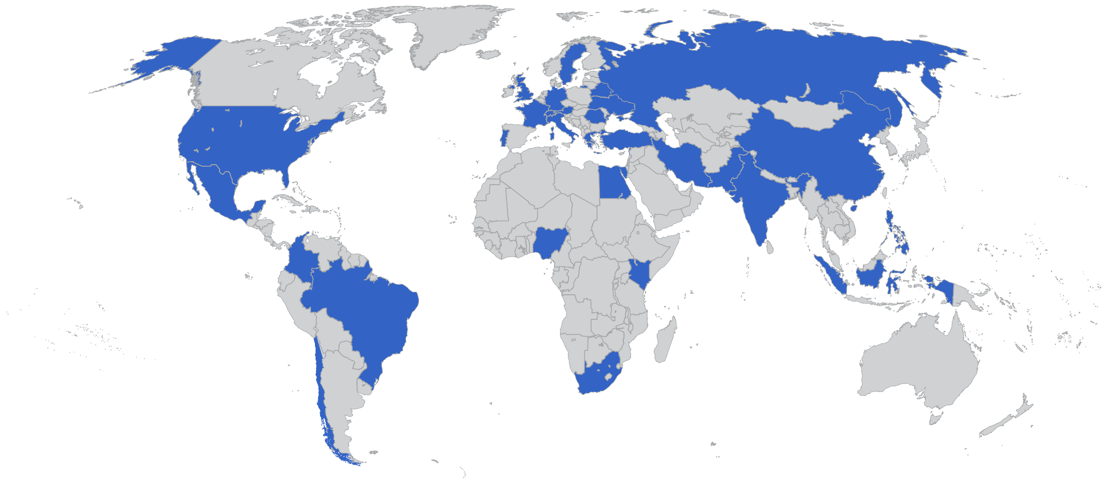


Fig.15: Countries of origin of the doctoral researchers of all six schools.

More detailed information on the progress of the six Information & Data Science Schools is provided in Chapter IV.

III. HIDA ACTIVITIES

In 2020, HIDA started various groundbreaking activities to improve the recruitment and training of data science talent - e.g. by providing networking initiatives such as the Trainee Network as well as international exchange programs, organizing recruitment events and datathons and setting up initial data science courses. Following on an overview of the versatile activities of HIDA in 2020 can be found.

1. TRAINEE NETWORK

The Trainee Network is a highly attractive Helmholtz-wide exchange program and a key element of HIDA: The program financially supports one- to three-month-long research stays at other Helmholtz Centers for doctoral and postdoctoral researchers (trainees) whose research has a strong connection to (applied) information or data sciences. The participating trainees not only apply their own knowledge at other centers, but also learn from the methods and approaches of colleagues from other disciplines. In this way, they expand their research portfolio, form new networks, and strengthen their ability to conduct research in an interdisciplinary environment. With the Trainee Network, HIDA creates new forms of collaboration and exchange between scientists of the different Helmholtz Centers. The centers increase their visibility within the Helmholtz Association, get to know up-and-coming data science talent from other research programs, and receive impulses for their own research areas. In addition, the Trainee Network promotes the mutual transfer of expertise in the field of Information & Data Science in the long term and ensures that methods and algorithms are shared between groups from different Helmholtz Centers and research domains. When fully established, more than 100 young scientists per year will be able to participate in the Trainee Network. In 2020, the first call for applications for the Trainee Network took place. Despite the pandemic situation, a total of 26 participants - doctoral researchers (15) and postdocs (11) - were selected to participate in the program and to visit another Helmholtz Center for a short-term research stay until the end of 2021. Nine of the selected doctoral researchers had no direct affiliation with one of the Helmholtz Information & Data Science Schools (HIDSS). Moreover, fourteen Helmholtz Centers will have participated in the exchange program as either sending and/or hosting institution by the end of 2021. Further reflection on the Association's six research fields highlights that a total of eleven exchanges will take place across borders of these research fields, e.g. from Health to Key Technologies/Information.

In summary, the participation of almost all centers in the 1st round as well as the high number of participants without direct HIDSS affiliation show that the HIDA Trainee Network will make a significant contribution to the Helmholtz-wide networking via smart minds – thus, leading to a direct benefit for all centers of the Association.

HIDA TRAINEE NETWORK – 1ST CALL FOR APPLICATIONS

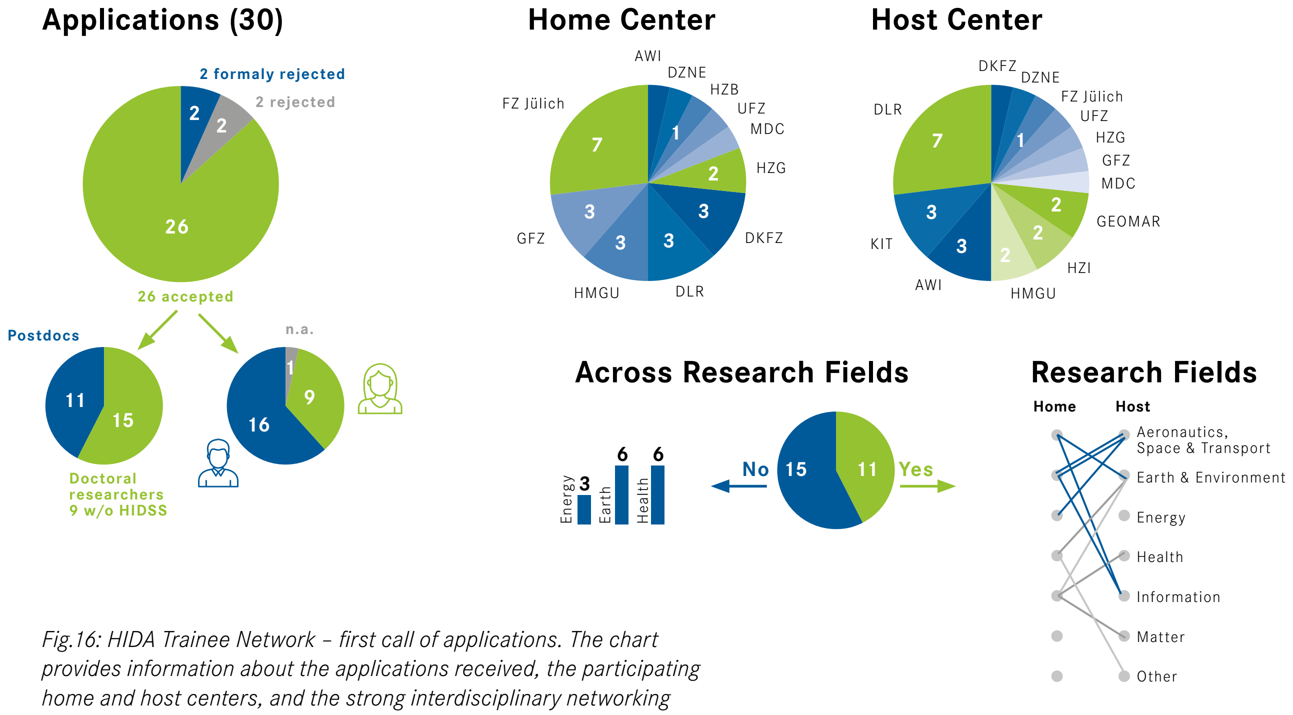


Fig.16: HIDA Trainee Network – first call of applications. The chart provides information about the applications received, the participating home and host centers, and the strong interdisciplinary networking that has been established through cross-disciplinary exchange.

2. INTERNATIONAL EXCHANGE PROGRAMS

HIDA aims to attract data scientists internationally and make them aware of career opportunities at all Helmholtz Centers. Exchange programs with particularly interesting partners are one way to engage talent from abroad and introduce them to Helmholtz, as well as foster scientific collaboration across borders at the same time. By presenting the vast opportunities all Helmholtz Centers and programs offer combined, HIDA maximizes the impact and efficiency of recruiting activities abroad, always promoting the brand of Helmholtz.

Israel is of strategic interest to the Association in this regard with influential applied data science research conducted at the country's research institutes as well as in its flourishing startup scene. Therefore, HIDA conducted a first pilot exchange program with Israel in 2020. Its first partner was the Data Science Research Center at Ben-Gurion University of the Negev (BGU). Since travel was not possible due to the COVID-19 pandemic, the exchange partners worked together on their projects online. Two data scientists from BGU worked with David Greenberg, head of Helmholtz AI's Young Investigator Group at HZG, on machine learning methods to quantify uncertainty in climate and weather models. Three data science PhDs and postdocs from the FZJ and KIT worked with Dr. Rami Puzis, Senior Lecturer at Ben-Gurion University of the Negev, on the topic of complex networks. The scientists expect to publish papers on these projects in 2021 – in some cases their first international publications and very first publications overall in their young careers.
In the forthcoming years, HIDA will pursue the initiation of further international exchange programs, starting with Canada and the US as well as EU partner countries. What's more, HIDA will grow the number of involved partner institutions by partnering with the Israel Data Science Initiative (IDSI). The Israeli Council for Higher Education established this initiative as part of a national funding program for data science. IDSI is a network for synchronizing activities among Israeli research institutions and for sharing international collaborations and cooperation with industry and public institutions.

3. FRIENDS OF HIDA

The long-term vision for the Helmholtz Information & Data Science Academy is to become a hub for data science education and training in Germany within a worldwide network united in its pursuit to contribute to solving grand challenges facing society, science and the economy. Together with the Helmholtz Information & Data Science Schools and excellent German university partners, HIDA welcomes academic and industry partners internationally to join us in this pursuit as partners in the Friends of HIDA network.

The network's purpose is to work with partners that add value to and work synergistically with the Helmholtz Centers, either in the education and further education of their data scientists, or in drawing attention to the Helmholtz Centers as an attractive employer in the Information & Data Sciences to attract talent. On a day-to-day basis, the partners share information about ongoing activities that are open to each other's data science communities with each other. They also plan and carry out joint activities with mutual benefits with each other.

HIDA founded the network in late 2019 and has since grown it to the following six partners:

- Alexander von Humboldt Institute for Internet and Society (HIIG)
- Data Science Research Center at Ben-Gurion University of the Negev (DSRC@BGU)
- German Academic Exchange Service (DAAD)
- Y-DATA - Yandex School of Data Analysis (YSDA)
- Weizmann Artificial Intelligence Center (WAIC)

Joint activities in 2020 included:

- With HIIG: A joint event targeted to parliamentarians to introduce them to AI education and research at Helmholtz. The event was postponed to March 2021 and will then be realized as a virtual event.
- With Akademie für Theater und Digitalität am Theater Dortmund: An exchange program for young data scientists and artists to explore and communicate data science in an artistic form at Helmholtz Centers and at the Akademie für Theater und Digitalität was planned in 2020 for 2021. This pilot exchange program will offer data scientists, artists and technicians the opportunity to conduct research at both the Academy and a Helmholtz Center in order to develop formats that make the topic of data science artistically accessible to the public and thus contribute to the transfer of science. The aim is to make the enormous possibilities, inherent in cutting-edge digital research, accessible to the theater of tomorrow. The call for proposals will open in the first quarter of 2021.
- With DAAD: In 2020, the DAAD had planned to take an international group of postdocs working on AI around Germany and presenting them with German AI institutions. HIDA had planned for the DLR and HGMU to be stops on the tour. Instead, online talks took place.

HIDA plans to grow the network following the Helmholtz data science communities' strategic needs and welcomes their recommendations on potential partners.

With this badge HIDA welcomes all new friends to its network.

4. FURTHER EDUCATION AND TRAINING

HIDA is committed to provide and promote training and further education for scientists in the field of Information & Data Science and strengthening the exchange of methods and knowledge transfer between the Helmholtz Centers. It also supports the other Helmholtz Incubator platforms and makes their expertise available to doctoral researchers of the Helmholtz Association working in this field. This means supporting local initiatives and activities, within graduate programs or institutes at Helmholtz Centers, to disseminate knowledge of methods and technologies of Information & Data Science throughout the Helmholtz Association as well as to share or develop foundational and/or advanced training offers.

4.1 HELMHOLTZ VIRTUAL ML SUMMER SCHOOL 2020

The development of the program for the first *Helmholtz Virtual ML Summer School* which took place from 21st September to 2nd October 2020 was a prototype and response to the external circumstances determined by the pandemic.

As HIDA aims to support the Helmholtz Centers and Helmholtz graduate programs to expand their local activities in the field of Information & Data Science thus enabling young scientist from all Helmholtz programs to participate. For this first prototype, HIDA collaborated with Helmholtz AI, the Munich School for Data Science (MUDS), the Institute of Statistical Learning and Data Science at Ludwig-Maximilians-Universität München (LMU), and the Munich Center for Machine Learning (MCML).

The core program of the *Helmholtz Virtual ML Summer School* 2020 was an introductory course in basic techniques and concepts of supervised machine learning “Introduction to Machine Learning (I2ML)”, which had been developed by researchers at LMU and a PI of the MUDS as a five days flipped classroom course and on site. As a virtual event, HIDA increased the original number of participants from 30 to 100 and since whole days were spent in videoconferencing HIDA stretched the program and made it 10 half days. The focus of the course was to provide a basic understanding of the various algorithms, models and concepts while explaining the necessary mathematical foundations.

The live sessions of the course were enriched with contributions from the cooperating partners:

- A keynote lecture and discussion with the participants by Prof. Dr. Bernd Bischl (Principal Investigator at MUDS and Helmholtz AI Associate, Professor and Chair of Statistical Learning and Data Science at LMU; and one of the directors of MCML).
- Two showcases by Dr. Dominik Thalmeier (Helmholtz AI consultant) “Using anomaly detection to identify mutations that effect hearing behavior” and by Christian L. Müller (Helmholtz AI consultant) “Sparse predictive modeling of microbiome data” providing insight into current research projects and the application of ML methods.

HIDA received 300 applications via the HIDA website of which 100 applicants were selected. With regard to recruiting HIDA accepted 25% applications outside from Helmholtz (mostly universities). A certificate of attendance was granted for an attendance of eight out of ten days.

The virtual event not only attracted attention from Helmholtz members, but also proved to be a useful tool to make the Helmholtz Association visible to the international interested professional public, especially for early stage researchers.

Facts and figures

300 Applications	100 Participants
	From 13 out of 18 Helmholtz Centers
60% from within Helmholtz	50% Helmholtz doctoral researchers
20% from other German research Institutions	25% other Helmholtz researchers
20% from international research institutions	25% other (mostly from Universities)

4.2 HIDA COURSE CATALOG

To gain visibility to the broad spectrum of already existing education offers in the field of Information & Data Science and disseminate these offers throughout the Helmholtz Association, HIDA started implementing an online course catalog that will be completed in 2021.

In 2020, HIDA started gathering sources of open education offers and networking events within the Helmholtz Association such as seminars, workshops, block courses, lectures, summer schools, datathons, meet-ups, symposia, conferences in an event calendar on the HIDA website[1].

Offers show a wide range of Information & Data Science topics from application of data science methods, data management to data ethics and sustainability of data or data-based knowledge production. They display profiles of Helmholtz Centers as well as association-wide interdisciplinary initiatives represented through the platforms of the Incubator or Helmholtz research fields.

Since it is open to public, the calendar makes Information & Data Science education offers visible and grants members of the Association an easy access to training. In making Helmholtz Centers and their programs visible, the calendar also supports HIDA’s recruiting efforts in the wider community.

The further development of the event calendar into a course catalog will focus on two user groups, organizers and attendees, and it will thus be designed to serve organizers of trainings to easily announce and disseminate their offers association-wide and provide easy access to the registration for attendees.

4.3 HIDA COURSE FUNDING

With HIDA Course Funding HIDA provides financial support for the organization and expansion of training offers and other formats that foster networking within the Helmholtz Association in the field of Information & Data Science and build up their research staff’s competencies. To enable local initiatives to share, open and expand their activities association-wide, HIDA set up administrative processes and guidelines to apply for financial support.[2] These guidelines were published in Q3 of 2020.

All members of the Helmholtz Association, especially Helmholtz graduate programs such as the Helmholtz Information & Data Science Schools can submit proposals for joint projects anytime, on the condition that these offers generate benefit and added value in the field of Information & Data Science not only locally but for the entire Helmholtz Association.

In the event of approval and building on the co-financing principle HIDA partially covers expense allowances (e.g. travel and accommodation costs) and remuneration for lecturers/trainers; services necessary for the event’s organization and promotion and to increase its outreach; rental costs for event venues.

HIDA received and approved proposals from AWI, DLR, HMGU, MDC but due to the COVID-19 pandemic plans for 2020 had to be cancelled resp. postponed to 2021.

Building on the HIDA Course Funding and contextual inquiries conducted with representatives from different stakeholder groups throughout the association, HIDA drafted a concept to provide Information & Data Science courses at all centers. The concept builds on a partnership with *The Carpentries* and the participation of all platforms.

The Carpentries is a non-profit internationally operating project that builds global capacity in essential data and computational skills for conducting efficient, open, and reproducible research. *The Carpentries* trains and fosters an active, inclusive, diverse community of learners and instructors that promotes and models the importance of software and data in research.

With this initiative, HIDA responds to demands on various levels and builds foundations to

- train trainers at all centers
- provide foundational and advanced training offers in Information & Data Science methods to develop reproducible and sustainable workshops tailored to local needs
- establish and provide best practices in a Helmholtz-wide network
- serve personnel developers at every Helmholtz Center to choose from a broad portfolio of education offers within the Helmholtz Association and the whole community of *The Carpentries*
- promote knowledge exchange networking and knowledge exchange at all levels between Helmholtz Centers, all Incubator platforms and researchers and with the open and international community of *The Carpentries*

A working group involving representatives from the Helmholtz Incubator platforms was formed in Q4 2020 and is pursuing to implement the first activities in Q2 2021.

[1] https://www.helmholtz-hida.de/en/activities/events/

[2] https://www.helmholtz-hida.de/en/activities/hida-course-funding/

5. NETWORKING EVENTS

5.1 HIDA REMOTE CHALLENGE ON CLIMATE CHANGE

In April 2020, HIDA organized the Virtual Challenge on Climate Change. Originally, this event was planned as a face-to-face event, which could not take place in the intended setting due to COVID restrictions. Instead, HIDA offered a virtual challenge at short notice: 32 researchers from four countries took part and worked on a question from the HZG. The topic of the challenge was on climate: the participants were to help find out how solar and volcanic activities from ancient times are reflected by climate models.

The new, digital format of the Virtual Challenge turned out to be an effective and quickly implementable way to inspire data science talent internationally to work on grand societal challenges using data science methods at Helmholtz.

HIDA plans to repeat this format in 2021. The event turned out to be an efficient way to reach HIDA's target group – young data scientists at Helmholtz Centers and other research institutions, nationally and internationally – and make them aware of the research that is conducted within Helmholtz, increasing interest in career opportunities. HIDA aims to take on topics from other research fields and to carry out the event with different partners in 2021.

5.2 HIDA DATATHON FOR GRAND CHALLENGES ON CLIMATE CHANGE

The Datathon for Grand Challenges on Climate Change took place in November 2020. The event was originally planned as a face-to-face format for April 2020. It was then moved and finally successfully converted into a virtual format and was streamed live via the HIDA website in two sessions.

The aim of the Datathon was:

- to develop solutions to climate-relevant scientific questions from the centers using data science methods
- to recruit new talent and establish contacts between scientists from the Centers and young data scientists at the beginning of their careers
- to promote the Helmholtz Association as an attractive employer in the field of Information & Data Science

During the Datathon, 94 Data Scientists from 12 Helmholtz Centers and seven countries worked in 13 teams on five questions, which were contributed and supervised by scientists from DLR, HZG, GEOMAR and UFZ.

These challenges were:

- **Map local city climate from satellite data automatically**: Xiaoxiang Zhu and her team at DLR asked to come up with machine learning models that create reliable Local Climate Zone maps for urban climatologists from satellite imagery.
- **Show the flow of fish larvae in warming oceans in an interactive visualization**: Willi Rath and the Ocean Dynamics Team at GEOMAR needed help with investigating human-induced changes at the beginning of the food chain.
- **Finding water with cosmic rays**: Martin Schrön collects data on soil moisture by a neutron detector installed on a car; he and his team at UFZ challenged the participants to develop a self-improving computer model that reliably detects landscape features in a set of images they have collected on their rides through Germany.
- **Spot the mistake in 50 million data points, cleverly**: Lennart Schmidt from the team around Corinna Rebman at UFZ sought help to find a clever method to automatically flag suspicious and bad data points in soil-moisture measurements in the German forest area Hohes Holz.
- **Developing reliable forecasts for drought**: Working with simulations spanning several thousands of years and helping Eduardo Zorita and his team from the Helmholtz Zentrum Geesthacht to develop a model that reliably predicts rain and snow for the following fall and winter season was challenge number five.

Set-up of the live stream at Haus Ungarn, Berlin.

The challenge participants were supported by HIDA partners: As technology partners, the Jülich Supercomputing Center and the Steinbuch Center for Computing of KIT provided HPC systems and computing time as well as support by one employee each. External industry partners of the event were Deloitte and NVIDIA. In addition to seven experienced data scientists from the centers, employees of the companies supported the participants as mentors.

Ten solution videos were submitted in the required time and a jury consisting of scientists from the centers that had submitted questions and two external jurors reviewed the submissions. The external jury members were: Prof. Dr. Hanna Meyer, professor and head of the Remote Sensing and Spatial Modelling Research Group at the Institute of Landscape Ecology at the University of Münster, and Dr. Marcel Dickow, head of the Federal Environment Agency's Digitization and Environmental Protection, E-Government unit (Umweltbundesamt).

In the end, five teams convinced the jury with their presentation. Juror Dr. Marcel Dickow was particularly impressed by the participants' ability to reflect on the specifics of the challenges they faced: "I think that's something what we need: That people are not only able to find solutions but also to understand the problem and to adapt their solutions to the problem."

It is a declared goal of the Datathon that this ability to reflect should actually contribute to solving the respective research questions. Challenge giver Eduardo Zorita is confident about the results found in his challenge to predict droughts: "All three teams did an excellent work within the tight timeframe, exploring different options. One of these solutions based on Random Forest did show promising results. They will be now further pursued and very likely included in our ongoing HAICU pilot project." Martin Schrön is also surprised about the good results: "It has given our idea a real push." He plans to invite the two winning groups to talk with them about a possible automation of the script.

HIDA is planning to continue this format in different scientific fields and with other and more partners in 2021.

5.3 HIDA @ RE:PUBLICA REMOTE

HIDA successfully placed a digital panel at the online conference *re:publica remote* on 7 May 2020. In the session, HIDA discussed with Prof. Dr. Alice McHardy (HZI) and Prof. Dr. Fabian Theis (HMGU) the topic *Data, data, data: What data helps in the fight against the virus?* The talk was streamed on re-publica.tv, Youtube and Facebook and communicated on all re:publica channels (social media, websites, newsletters) as well as via the participating research centers. The contribution ranked among the top 4 most-watched sessions. With this contribution the panel raised awareness of the Helmholtz Association and the growing importance of data science for science. It also highlighted the importance of training and recruiting young talent in this field. The video of the session is available on the HIDA website. In 2021, HIDA plans a further virtual conference participation.

5.4 HELMHOLTZ VIRTUAL DATA SCIENCE CAREER DAY

The Helmholtz Virtual Data Science Career Day took place for the first time on 23 September 2020. The pioneering event attracted 547 visitors from over 80 countries. Overall, Helmholtz as an attractive employer brand received a lot of international attention with ads reaching 1.5 million impressions. The event page achieved over 4,000 clicks and over 2,300 registrations from 130 countries for the Career Day. At the 23 online exhibition stands, employees of the following Helmholtz Centers, Helmholtz Information & Data Science Schools and HIDA partners were in contact with interested parties in about 400 chats:

Helmholtz Centers:

- HZG, UFZ, DESY, DLR, FZJ, DKFZ, DZNE, MDC, KIT, GFZ, CISPA

Schools:

- *HDS-LEE* – Helmholtz School for Data Science in Life Earth and Energy
- *HIDSS4Health* – Helmholtz Information & Data Science School for Health
- *MUDS* – Munich School for Data Science
- *DASHH* – Data Science in Hamburg Helmholtz Graduate School for the Structure of Matter
- *MarDATA* – Helmholtz School for Marine Data Science

HIDA partners:

- DAAD
- Alexander von Humboldt Institut für Internet und Gesellschaft
- Data Science Research Center @ Ben-Gurion University of the Negev
- Yandex School of Data Science
- CASUS Center for Advanced Systems Understanding

Helmholtz Incubator Platforms:

- Helmholtz AI, HIP Helmholtz Imaging Platform

The HIDA Data Science Job Board, which provides information on all open Data Science positions in the Helmholtz Association, had almost 600 page views on this day alone. For further information on the Data Science Job Board please see *page 23*.

Live virtual talks and panel discussions enriched the program. The conference program with three keynote speeches and four panels provided information on career development and areas of work at Helmholtz. The above figures indicate that the event lived up to its goal of making data scientists aware of the interdisciplinary career opportunities at Helmholtz.

According to a post-event survey, exhibitors and partners were equally happy with the event's outcome and the majority would be interested in partaking in another Helmholtz Virtual Data Science Career Day in the future. Therefore, HIDA plans to repeat and expand the Helmholtz Virtual Data Science Career Day in 2021.

The event in numbers:

Registration	2339
Visitors	547
Visitors' countries of origin	80
Chats	399
Average number of booth visits	585

HiDA HELMHOLTZ Information & Data Science Academy
QUESTIONS? LET'S CHAT!
Visit our courses and events
Find your next job
SKILLS FOR THE FUTURE OF RESEARCH
COLLABORATE & BUILD YOUR NETWORK
LEARN MORE ABOUT OUR DOCTORAL RESEARCHERS
DISCOVER OUR INFORMATION & DATA SCIENCE SCHOOLS
MarDATA
HDS-LEE
DASHH
HEIBRiDS

The HIDA booth during the Helmholtz Virtual Data Science Career Day.

6. OUTREACH

6.1 HIDA WEBSITE

In 2020, HIDA has greatly expanded its website in order to make HIDA's numerous new activities visible and to continue developing the website into a platform that promotes exchange programs, courses, events, research projects, and jobs in the field of Information & Data Science at Helmholtz both nationally and internationally.

To this end, HIDA has greatly expanded its news section with portraits of Helmholtz scientists and news from the field of Information & Data Science. In total, the news section has grown to 51 articles - including 12 portraits presenting the research projects of doctoral researchers from the Helmholtz Information & Data Science Schools.

Since December 2020, selected articles have been presented in a Helmholtz Teaser Box on Spektrum.de to further increase coverage. In addition, HIDA has established content partnerships with magazines of various Helmholtz Centers: Among them the magazines of DLR, KIT, FZ Jülich and UFZ. In addition, some of HIDA's news articles are regularly republished on Helmholtz.de and on the website of the Helmholtz Climate Initiative.

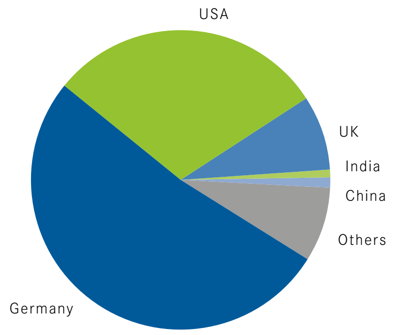
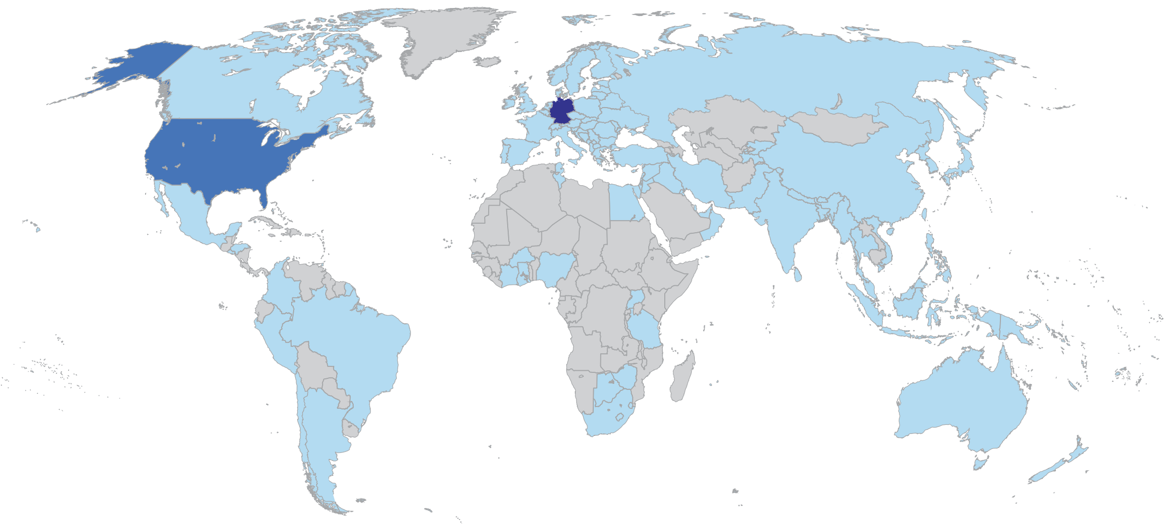


Fig. 17 & 18: Visitors came from the following countries: 52% Germany, 30% USA, 8% UK, 1% India, 1% China, 8% other countries.

With the website up and running, traffic on the new HIDA website in 2020 has also already increased noticeably. In total, the HIDA website recorded 77,363 page views from over 87 countries in 2020. Visitors came from the following countries: 52% Germany, 30% USA, 8% UK, 1% India, 1% China, 8% other countries. Numerous visitors were routed from the HIDA page to the page of a Helmholtz center or school.

41% of visitors accessed multiple pages on the HIDA website. The largest share of traffic went to HIDA's offerings and HIDA's Data Science Job Board, thus heavily promoting the recruiting activities of all centers participating.

Visitors map



6.2 HIDA DATA SCIENCE JOB BOARD

With the HIDA Data Science Job Board, HIDA aims to make data science talent who visit the HIDA website aware of data science positions at Helmholtz Centers. Throughout 2020, 60-80 Data Science-related positions within the Helmholtz community have been open at all times. HIDA collected and posted them on their website as a searchable database. In total, the HIDA Data Science Job Board recorded over 11,000 views in 2020. Much of the traffic on the job board resulted in calls to the detailed job postings on the centers' websites. The overview of the transitions in 2020 also clearly shows this:

Top 12 transitions from HIDA website to job-/websites of Helmholtz Centers & Schools:

1. www.dlr.de: 794 additional visits
2. www.fz-juelich.de: 745 additional visits
3. www.hds-lee.de: 356 additional visits
4. www.desy.de: 298 additional visits
5. www.mdc-berlin.de: 303 additional visits
6. www.heibrids.berlin: 301 additional visits
7. www.mu-ds.de: 294 additional visits
8. gfz-potsdam.concludis.de: 266 additional visits
9. kit-stellenangebote.dvinci-easy.com: 253 additional visits
10. www.hidss4health.de: 248 additional visits
11. www.dashh.org: 240 additional visits
12. www.hzdr.de: 187 additional visits

6.3 SOCIAL MEDIA

In order to make HIDA's activities and news, as well as the data science jobs at the Helmholtz Centers accessible to an even larger audience, HIDA greatly expanded its activities in the social media channels in 2020. HIDA's Twitter channel @HIDAdigital grew from approximately 400 to over 2,100 followers, with posts reaching over 1.2 million tweet impressions in 2020. The HIDA Twitter profile had over 16,700 views in 2020.

In addition, HIDA launched a newsletter in June 2020, which now has over 440 subscribers. Since the end of July 2020, HIDA has also been active with a profile of its own on LinkedIn, where it now has more than 600 followers. With its posts on LinkedIn, HIDA still reached over 35,500 impressions by the end of 2020.

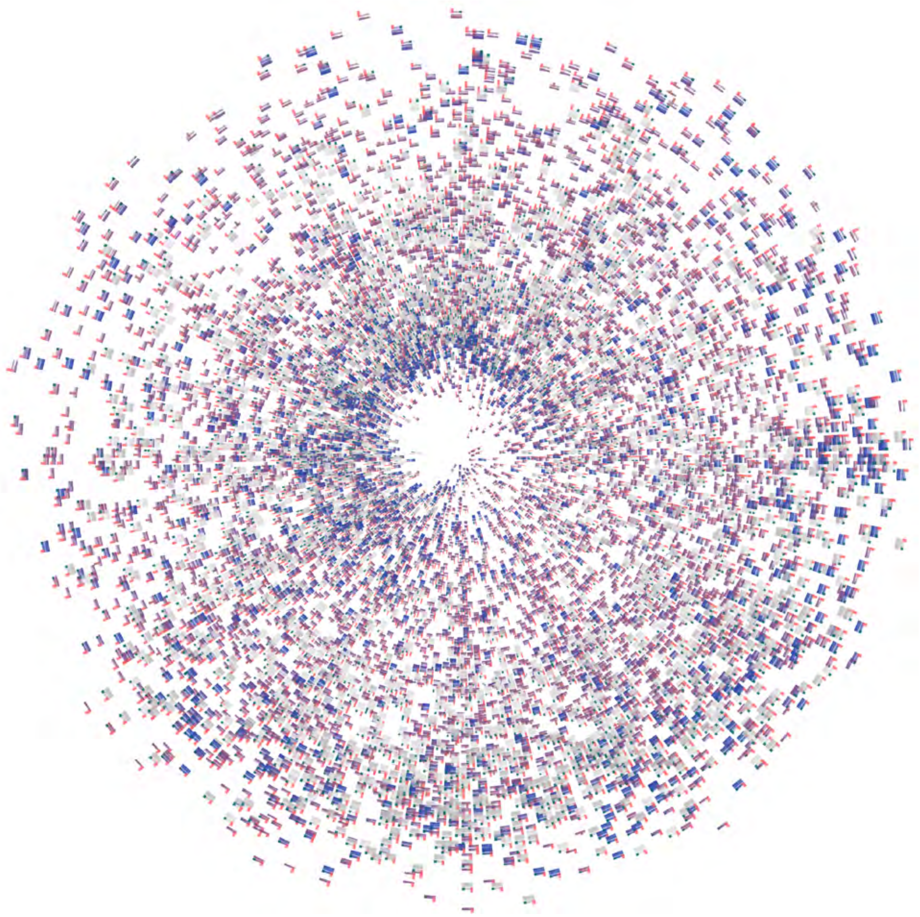
6.4 THE MAGIC OF DATA

Under the title "The Magic of Data", HIDA created artistic visualizations of exemplary data-driven research projects from five Helmholtz Information & Data Science Schools to communicate the importance of data-driven research to the public in a new and engaging way. With the project HIDA wanted to give the complex topic of data science a visual form that allowed people to experience it using their senses. The goal was to create a means of accessing modern, data-driven research via the images.

To do this, HIDA collected 12 datasets from doctoral researchers at the Helmholtz Information & Data Science Schools. Scientists use the datasets to study a range of topics in the research fields Energy, Earth and Environment, Health, Key Technologies, Matter, and Aeronautics, Space, and Transport:
e.g. the thawing dynamics of Arctic permafrost, biological control mechanisms that regulate our internal clock, machine learning in neurosciences and cell biology, x-ray models of electronic structures of substances, metabolic processes in microorganisms, precise self-localization systems, neural networks to determine ozone concentration, AI-aided evaluation of earthquakes, radiotherapy analysis and predicting enzyme structures.

For HIDA, the Cologne based designer Michael Schmitz from NEOANALOG translated the datasets into computer generated, aesthetically loaded visuals. The designer began his work by writing an algorithm that served as a set of rules and "translated" the data points – which initially only existed as numerical values in long Excel lists – into graphical elements such as coordinates, radii, line widths, and radians. In addition, chance played a role in this "translation" process, with the program selecting new colors and shapes to generate the respective image. This means that the results were no information graphics aiming to make the data understandable, but rather artistic illustrations of the essential subject the doctoral researchers at the Helmholtz Information & Data Science Schools are working with.

To make the visualizations accessible to a broad public, HIDA presented the impressive images and animations at the re:publica Campus 2020 in Berlin from September 06 - October 04 together with thorough information regarding the Helmholtz research programs from which the data originated. In addition, the HIDA website provides information about the research projects behind it. HIDA offers the visualizations on its website for free use to all centers.



One of the visualizations of data-driven research projects at the Helmholtz Data Science Schools.

7. HIDA WORKSPACE IN CENTRAL BERLIN

The furnishing of the HIDA workspace was completed in 2020. The HIDA event space of approximately 80 square meters offers the infrastructure for a wide range of event formats - from smaller meetings to larger events. All rooms come with modern technical equipment and access to fibre optic cable and are therefore an excellent hub where data-driven minds are able to come together in the heart of Berlin (Friedrichstraße 171, 10117 Berlin).

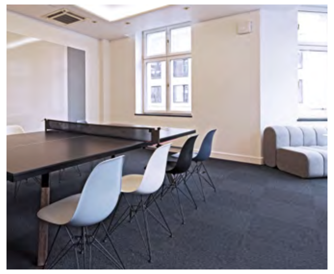
HIDA aims to promote collaboration and the interdisciplinary exchange of ideas in Information & Data Science. To this end, the HIDA event space is available to employees of all Helmholtz Centers, programs, working groups and committees who want to host data science events and meetings in order to bring together data scientists and to nurture the exchange of knowledge and data science methods.

THE LAB

This flexible event space offers the freedom to organize an event as desired - whether it is a lecture for a larger audience or a small meet-up. The Lab has space for up to 60 people. A partition is available in order to split the space into two rooms and adapt it to the needs of the respective event.

THE ARENA

The arena is the most flexible space. It combines a meeting space with room for catering with a kitchen unit. The modular elements of the room allow organizing the space to suit the occasion – from high-profile events such as receptions to informal work meetings in a relaxed atmosphere.



THE CREATIVE BOX

Guaranteed to be "out of the box": Thanks to its large magnetic wall, the Creative Box is ideal for creative workshops, presentations, and meetings for up to 12 participants, allowing them to work in small, focused groups.

From day one, the HIDA Workspace was used intensively from numerous internal and external data science groups. Due to the COVID pandemic, the HIDA workspace was only used for a very limited number of events throughout 2020. In 2020, the HIDA Office accommodated (only, due to COVID) 26 events with a total number of 360 participants. Amongst other, the following meetings took place:

- HIFIS Cloud Services Workshop
- Helmholtz Metadata Collaboration Platform (HMC) Kickoff
- Berlin Bayesians Meetup
- Meeting of the executive committee of the Helmholtz Association
- Awarding of Helmholtz-Ausbildungspreis 2020
- General Meeting of the de-RSE e.V. - Society for Research Software
- Meeting of the working group "Digital Qualified Staff" of the priority initiative "Digital Information" of the Alliance of German Science Organizations

The mission, stated in the original concept, to make the HIDA office an inspiring place to meet will be fully pursued again once the COVID restraints have fallen for good.

10. HIDA STEERING BOARD

The General Assembly of the Helmholtz Association has established the Steering Committee of the Helmholtz Information & Data Science Academy (HIDA-Steer). The following top-level researchers have been recruited for the steering, monitoring and strategic orientation of HIDA's activities:

- Sabine Attinger, Helmholtz Centre for Environmental Research – UFZ
- Stephan Frickenhaus, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI)
- Veit Hagenmeyer, Karlsruhe Institute of Technology (KIT)
- Uwe Konrad, Helmholtz-Zentrum Dresden-Rossendorf (HZDR)
- Klaus Maier-Hein, German Cancer Research Center (DKFZ)
- Alice McHardy, Helmholtz Center for Infection Research (HZI)
- Uwe Ohler, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC)
- Nina Rohringer, Deutsches Elektronen-Synchrotron (DESY)
- Achim Streit, Karlsruhe Institute of Technology
- Wolfgang Wiechert, Forschungszentrum Jülich
- Xiaoxiang Zhu, German Aerospace Center (DLR)

Two virtual HIDA-Steer Meetings took place (20.04.2020 and 26.08.2020).

Furthermore, in 2019, HIDA has established a contact person ("HIDA Liaison Officer") at each Centre in order to maximize the benefits of HIDA for all Helmholtz Centers. Through them, the Helmholtz Centers always receive all relevant information on planned HIDA activities directly.

In 2020, HIDA organized two virtual HIDA Liaison Officer and School Coordinator Meetings (22.04.2020 and 11.09.2020).

Wolfgang Wiechert
FZ Jülich

Veit Hagenmeyer
KIT

Sabine Attinger
UFZ

Uwe Konrad
HZDR

Stephan Frickenhaus
AWI

Alice McHardy
HZI

Uwe Ohler
MDC

Nina Rohringer
DESY

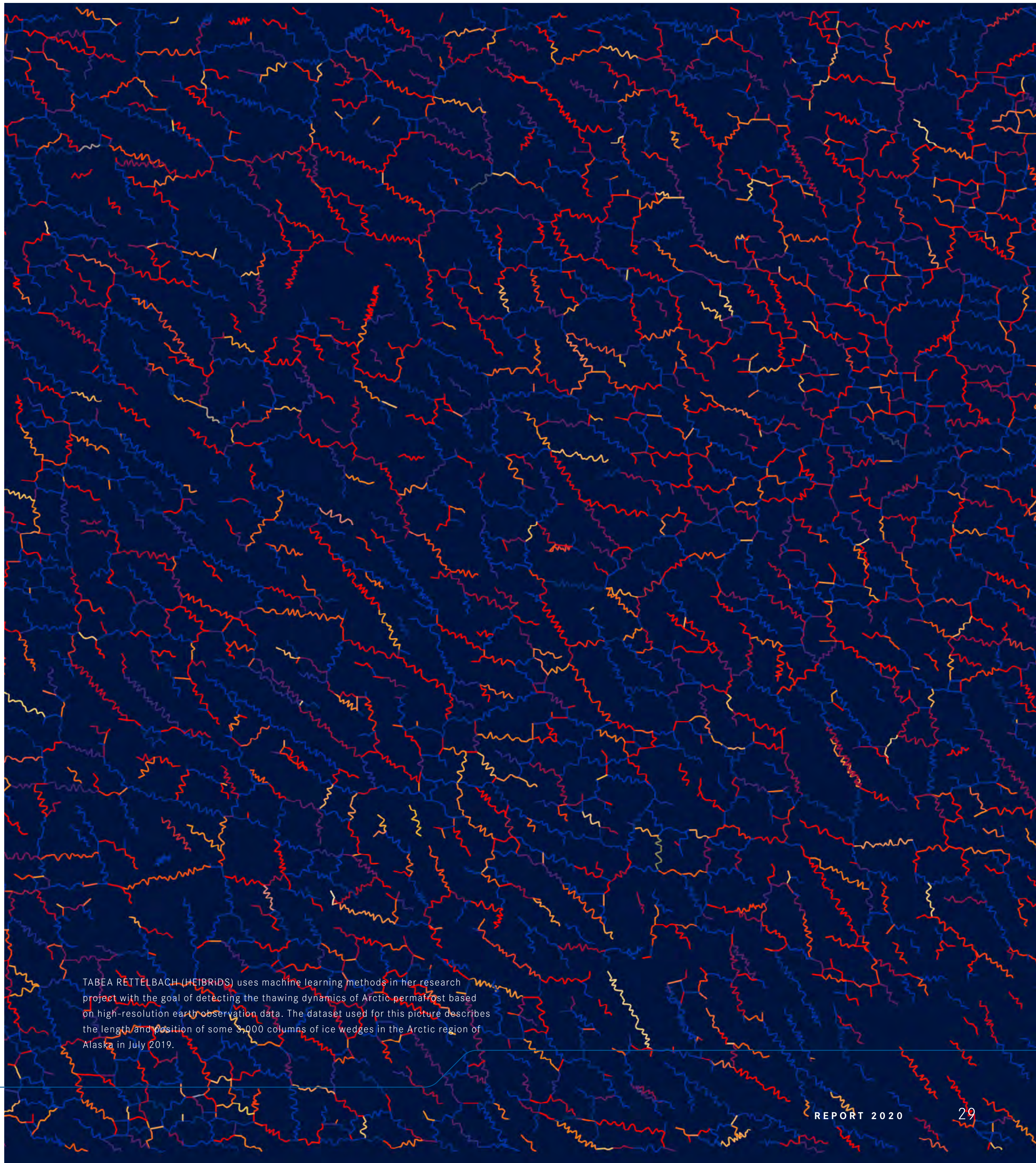
Xiaoxiang Zhu
DLR

Achim Streit
KIT

Klaus Maier-Hein
DKFZ

HEIBRIDS
HELMHOLTZ EINSTEIN INTERNATIONAL
BERLIN RESEARCH SCHOOL IN DATA SCIENCE

IV. FACTS & FIGURES ON THE HELMHOLTZ INFORMATION & DATA SCIENCE SCHOOLS



TABEA RETTELBACH (HEIBRiDS) uses machine learning methods in her research project with the goal of detecting the thawing dynamics of Arctic permafrost based on high-resolution earth observation data. The dataset used for this picture describes the length and position of some 3,000 columns of ice wedges in the Arctic region of Alaska in July 2019.

HEIBRIDS – HELMHOLTZ EINSTEIN INTERNATIONAL RESEARCH SCHOOL IN DATA SCIENCE

HEIBRiDS brings together six Helmholtz Centers and four university partners from the Einstein Center Digital Future (ECDF), which focuse on core digitization technologies, from digital health to digital industry and the digital humanities. The participating Helmholtz Centers have first-class expertise in the fields of molecular medicine, astrophysics, polar and marine research, aerospace, materials science and geosciences. The goal of HEIBRiDS is to train a new generation of researchers who are both qualified Data Scientists and understand the requirements and challenges of those disciplines in which Data Science has become indispensable.

Research areas:
The participants of the HEIBRiDS program are working in very different research areas, reaching from Earth & Environment, Astronomy, Space & Planetary Research to Geosciences, Materials & Energy and Molecular Medicine.

Partners of HEIBRiDS:
Alfred Wegener Institute, German Electron Synchrotron, German Aerospace Center, German Research Center for Geosciences, Helmholtz-Zentrum Berlin für Materialien und Energie, Max Delbrück Center for Molecular Medicine and the Einstein Center Digital Future in Berlin with Technical University, Charité University Medicine, Freie Universität Berlin and Humboldt University

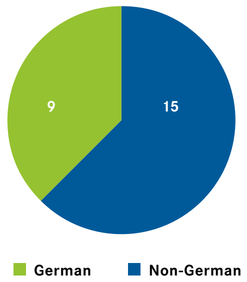


Fig. 23: Recruitment of doctoral researchers from Germany and abroad at HEIBRIDS 2019–2020 (excl. associated doctoral researchers).

Applicant situation and recruitment

A total of 24 positions have been occupied at HEIBRiDS (associated doctoral researchers not included): 13 PhD researchers started in the first round in 2018 (one of them became an associated doctoral researcher in 2020), further two positions were awarded in the second round 2019, followed by 10 PhD positions in 2020. The two candidates in 2019 were filtered out of 72 applications, of which only seven were from Germany. In 2020, 303 applications reached the school, 276 of them from international applicants.

Nine German PhD researchers and 15 from abroad are currently conducting research at HEIBRiDS (excl. associated doctoral researchers). Seven women are represented among the PhD candidates. In addition, there are two female associated researchers and five male associated researchers. These come from the following institutions: Max Delbrück Centrum (2), German Aerospace Center with TU Berlin (1), TU Berlin (1), Humboldt Universität zu Berlin (1), TU Dortmund (1) and Alfred Wegener Institute and Humboldt Universität zu Berlin (1).

.

Events and networking

Whereas in 2019 22 events took place at HEIBRiDS, the number increased in 2020 up to 31:

2nd HEIBRiDS Retreat, 29.-30.10.2020, Online event (25 participating Doctoral researchers, 5 associated students and 35 PIs): The schedule of the 2nd HEIBRiDS Retreat included two guest lectures, Pecha Kucha presentations from the first and second cohort of doctoral researchers, one-slide presentations from the third cohort of Doctoral researchers and Round Table discussions in small groups of PIs and Doctoral researchers on selected project challenges.

Orientation and Design Thinking Day for the 3rd cohort of HEIBRiDS students, ECDF & BIMSB/MDC, 07.10.2020, Online event (7 participating HEIBRiDS doctoral researchers and 3 associated doctoral researchers): On the first half of the day HEIBRiDS was introduced to the new cohort of doctoral researchers and answered their questions. During the second half of the day, HEIBRiDS ran a teambuilding activity based on the Design Thinking concept.

Workshop on Open Science, 25.11.2020, Online event, (15 participating doctoral researchers): Open Science concepts and tools have the potential to transform the current scientific system for the greater good of all. This interactive workshop was an introduction to what Open Science is and why it is needed. There was an overview of the main areas of Open Science: Open Access, Open Data, Public Engagement, practical tips, and engaging activities.

13 HEIBRiDS Lectures

- 08.01.2020, Petra Ritter (Charité Universitätsmedizin Berlin): “The Virtual Brain”
- 22.01.2020, Frederik Tilmann (GFZ): “Seismology for Earth Imaging in the Age of Large Distributed Datasets”
- 05.02.2020, Stefan Klus (FU): “Learning Dynamic Laws from Data for Complex Processes”
- 06.05.2020, Daniel D. Hromada (UdK): “From Text to Categories and Beyond: You don’t need “deep learning” to do it”
- 20.05.2020, Christoph Lippert (Hasso Plattner Institut): “Detecting Heritable Phenotypic Traits in Images”
- 03.06.2020, Ziawasch Abedjan (TU Berlin): “Example-Driven Data Cleaning”
- 01.07.2020, Norbert Marwarn (PIK Potsdam): “Methods for Investigating Recurrence Phenomena in the Earth System”
- 21.10.2020, Luca M Ghiringhelli (Fritz Haber Institute of the Max Planck Society): “Beyond Just Fitting Numbers: Interpretable Artificial Intelligence for the Small Data of Materials Science”
- 04.11.2020, Sayan Mukherjee (Duke University): “Modeling Time Series and Dynamical Systems”
- 18.11.2020, Anders Gorm Pedersen (Technical University of Denmark): “Bayesian Inference – an Introduction and Examples of its Application to Bioinformatics Research”
- 02.12.2020, Anna Bauer-Mehren (Roche Pharma): “Unlocking the Potential of FAIR Data Using AI at Roche Pharma R&D”
- 09.12.2020, Johann-Christoph Freytag (HU/ECDF): “Scalable Processing of Scientific Data in the Age of Data Science”
- 16.12.2020, Stephan Frickenhaus (AWI): “On the Creation of Virtual Research Environments for Data Analytics”

15 PhD seminars

Publications

A total of 13 first author publications (journal/conference – 9 submitted, 4 published) were finished as well as 6 co-authored publications (journal/conference – 4 submitted, 2 published).

First author (published/peer-reviewed):

1. P. Tillmann, K. Jäger and C. Becker (2020), Minimising the Levelised Cost of Electricity for Bifacial Solar Panel Arrays Using Bayesian Optimization, Sustainable Energy Fuels, 4, 254-264. 10.1039/C9SE00750D
2. S. Agarwal, N. Tosi, D. Breuer, S. Padovan, P. Kessel and G. Montavon (2020), A Machine-Learning-Based Surrogate Model of Mars' Thermal Evolution, Geophysical Journal International, 222(3), 1656-1670. https://doi.org/10.1093/gji/ggaa234
3. L. Weber, K. Thobe, O.A.M. Lozano, J. Wolf and U. Leser (2020), PEDL: Extracting Protein-Protein Associations Using Deep Language Models and Distant Supervision, Int. Conf. on Intelligent Systems in Molecular Biology, 36, Supp.1, i490-i498. https://doi.org/10.1093/bioinformatics/btaa430
4. J. Münchmeyer, D. Bindi, U. Leser and F. Tilmann (2020), The Transformer Earthquake Alerting Model: A New Versatile Approach to Earthquake Ear-LY Warning, Geophysical Journal International, ggaa609. doi.org/10.1093/gji/ggaa609

First author (submitted/peer-reviewed): **missing info for 7 submissions**

1. J. Münchmeyer, D. Bindi, U. Leser and F. Tilmann (2021). Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network. https://arxiv.org/abs/2101.02010
2. A.H.C. Vlot, S. Maghsudi and U. Ohler (2020). SEMITONES: Single-cEll Marker IdentificaTiON by Enrichment Scoring. bioRxiv 2020.11.17.386664. https://doi.org/10.1101/2020.11.17.386664

Co-authored (published/peer-reviewed):

1. J. Ren, L. Lin, K. Lieutenant, C. Schulz, D. Wong, T. Gimm, A. Bande, X. Wang and T. Petit (2020). Role of Dopants on the Local Electronic Structure of Polymeric Carbon Nitride Photocatalysts. Small Methods 2000707. https://doi.org/10.1002/smtd.202000707
2. K. Jäger, P. Tillmann, E.A. Katz, and C. Becker (2020) Perovskite/Silicon Tandem Solar Cells: Effect of Luminescent Coupling and Bifaciality. Sol. RRL. https://doi.org/10.1002/solr.202000628

Co-authored (submitted/peer-reviewed): **missing info for 3 submissions**

1. R. Shahan, CW. Hsu, T.M. Nolan, B.J. Cole, I.W. Taylor, A.H.C. Vlot, P.N. Benfey and U. Ohler (2020), A Single Cell Arabidopsis Root Atlas Reveals Developmental Trajectories in Wild Type and Cell Identity Mutants, bioRxiv 2020.06.29.178863. https://doi.org/10.1101/2020.06.29.178863

Further News from HEIBRiDS

- The School was present at the following conferences or events:
- 07.05.2020, EGU General Assembly, Vienna, Austria, oral presentation
- 04.-08.05.2020, EGU virtual conference, oral presentation
- 16.-19.09.2020, 13th annual RECOMB/ISCB Conference on Regulatory & Systems Genomics with DREAM Challenges, poster presentation
- 20.-22.10.2020, Taiwan Earthquake Research Center Annual Meeting, 2 poster presentations
- 23.11.2020, 73rd Annual Meeting of the APS Division of Fluid Dynamics, online, oral presentation

Communication/Marketing

HEIBRiDS has a **website** (www.heibrids.berlin), a **monthly newsletter** is sent to all students and PIs.

GREGOR PFALZ – HEIBRIDS

Traveling through Time in the Arctic

Gregor Pfalz has the air of a young researcher who is burning to get to the bottom of things. Which fits perfectly, because his work as a doctoral researcher at the HEIBRiDS Data Science School in Berlin involves analyzing data from sediment collected from Arctic lakes, with the aim of making predictions about the climate of the future. And how does this work? It takes a lot of patience and a merging of two disciplines - geology and computer science

A few things have already happened before Gregor Pfalz gets hold of the data he wants to use to look into the future. For decades, researchers on expeditions to Arctic regions in Russia have driven snowmobiles onto frozen lakes, drilled holes in the ice sheets, set up tripods, and hammered metal tubes into the sediment at varying depths in the lakes, or taken boats out in the summer to collect samples - of the fine, silty clay at the bottom of the water, which holds the key to the past.

Researchers are still taking these samples to this day. "Sediment like this is a great climate archive," explains Pfalz, his ice-blue eyes sparkling with enthusiasm. After all, it contains information that can be used to reconstruct what the climatic conditions of an ecosystem were like in the past - similar to the insights provided by tree rings. Researchers speak almost lovingly of "climate proxies", including minerals, carbons, and diatoms, that can be used to draw conclusions about metabolic processes in the past - and therefore in the future as well.

In other words, sediment samples essentially serve as history books for geologists. And the rule of thumb is that the longer the samples are - or the deeper researchers drill into the lakebed - the better. "Some of them let us look as far back as 50,000 to 60,000 years into the past," says Pfalz, who wants to do exactly this. The doctoral researcher at the HEIBRiDS graduate school is writing his thesis in the field of paleoclimatology and applying data science methods to this end. His research project is titled "Arctic Environmental Data Analytics."

"We use lake sediment to look into the past to make predictions for the future"

At a workshop, Pfalz was once asked to sum up his research in a sentence. He laughs, because his colleagues at the workshop thought his answer sounded like a haiku: "We use lake sediment to look into the past to make predictions for the future."

To get to the bottom of how Pfalz ended up in this field of research, we first dig into his own history: What is his background, his "sediment"? Pfalz was born in 1989 in Bautzen, Saxony, a German town known for its mustard and the home of pop rock band Silbermond. He also finished school in Bautzen and did a year of voluntary cultural work there before going to Dresden to study water management with the aim of becoming an engineer.

Then two things happened that prompted him to make the move to Potsdam. First, a professor from the US told him they had a good research program there. Second, Pfalz was doing an internship at the EAWAG institute in Switzerland when Antje Boetius, Director of the Alfred Wegener Institute for Polar and Marine Research (AWI) gave a talk there. He was excited by what he heard and learned soon after that AWI had a place open for a Master's candidate in Potsdam. Pfalz applied, got the place, and wrote his thesis about the erosion of Arctic coasts in Canada.

In other words, Pfalz already knew a fair amount about the Arctic. But after his Master's he ran into a bit of an ice age in his career - which actually proved to be very productive later on: He wanted to do his doctorate but was not able to work with the professors he had been hoping for. So, he used this time to hone his skills and actually learned the programming language Python - which made him realize how much he enjoyed computer science.

Following this discovery, Pfalz stumbled across the HEIBRiDS Data Science School in Berlin - an interdisciplinary graduate program founded by Helmholtz and the Einstein Center Digital Future (ECDF) in 2018. The program aims to train a new generation of researchers who are both skilled data scientists and work in scientific disciplines where progressive approaches to handling data are becoming increasingly important. Geology is one such discipline. And AWI happens to be one of the partners HEIBRiDS works with.

Once again, Pfalz applied, and was once again offered the position. He has now been working on his doctoral thesis since 2018 in his office at AWI, which is on Telegrafenberg hill in Potsdam, and has two supervisors here - both geologists- and another at Humboldt University who is a professor of computer science.

At HEIBRiDS, Pfalz comes into contact with like-minded people; the doctoral researchers he meets here are also working at the interface between their particular field of research and the data sciences, and he is able to discuss ideas and findings with them. He likes this and the graduate school's concept of giving its doctoral researchers time to learn about whichever domain is less familiar to them so they can work on a truly interdisciplinary basis. For example, Pfalz has taken further database and machine learning courses at the universities partnering with HEIBRiDS. He has also attended numerous talks at the ECDF given by researchers in fields ranging from the geosciences and biology to medicine, who discussed how the data sciences can be applied in practice.

"I try to look at my topic from a data sciences perspective at every stage, too"

"All these aspects of the program let us look at the even bigger picture and truly think outside the box," says Pfalz, who is clearly bursting with a young scientist's enthusiasm for his research project. "That's why I try to look at my topic from a data sciences perspective at every stage, too," he says. "And use methods that maybe haven't been used by anyone else before."

Pfalz notes that the Russian part of the Arctic has often been underestimated in climate research prior to this. But the data that come from this region are important for climate modelers who want to understand the impact certain processes have at the global level. If Russian regions of the Arctic are not taken into account, they will be missing crucial insights. That is why Pfalz is taking another closer look at the data now and working with data science methods to try and detect patterns or pinpoint turning points in the past. These are of special significance for climate research because they allow predictions to be made about the future later on as well. These could basically be described as if-then conclusions.

In his first two years as a doctoral researcher, Pfalz focused on how to enable comparisons between data from expeditions in past decades - which were also collected using different methods in some cases. The Arctic also presents a further challenge, as the field of maritime research distinguishes between varved and unvarved lakes. In varved lakes, the layers of sediment that permit conclusions to be drawn about how a lake changed over the course of a year are clearly visible; this is not the case with unvarved lakes. Lakes in the Arctic are primarily unvarved, which means Pfalz has to constantly revisit his data. He continually draws on his expertise in the geosciences to understand these data, and his knowledge of the data sciences is crucial to discovering patterns therein. "It's the interplay between the two disciplines that helps me," says Pfalz. And the patience to take another look, over and over again.

But unlike the fields of many of the other doctoral researchers at HEIBRiDS, Pfalz is not working with big data in his project. Rather, he looks at numerous small datasets and tries to glean as much information as possible from them. "We have to work with a lot of uncertainty in the field of paleoclimatology," he says. "I'm trying to improve this situation so we're able to make truly clear statements." After all, this is Pfalz's objective, and is what drives him: He wants to gain insights that enhance analytical capabilities and, in turn, enable climate modelers to make better predictions about the future. And these predictions will ultimately result in better decisions - for the climate of the future.

Author: Andrea Walter

https://www.helmholtz-hida.de/en/activities/news/newsdetail/traveling-through-the-arctic/

Credits: Gregor Pfalz: private/Arctic ice: Alexander Pogorelsky/Unsplash

OLGA KONDRATEVA – HEIBRIDS

Helping Satellites Connect

To spot fires and predict droughts, satellites need to communicate efficiently with Earth – and data science algorithms can help. Olga Kondrateva, doctoral researcher at the Helmholtz Einstein International Berlin School in Data Science (HEIBRiDS), is programming the satellites to pick and choose what data is worth transmitting

Olga Kondrateva has always been interested in efficient communication.

As an undergraduate in St. Petersburg, Russia, Kondrateva studied applied linguistics, delving deep into the structures that make languages work. Today, she's a doctoral researcher at the Helmholtz Einstein International Berlin School in Data Science (HEIBRiDS), using her programming skills to help satellites communicate better with researchers on Earth.

It's an important job. There are thousands of satellites ceaselessly circling the planet, many of them busy taking pictures and transmitting images back to analysts on the ground. In recent years, pictures taken from these "low -Earth orbit," or LEO, satellites have been used to guide responses to natural disasters and study the effects of climate change.

LEO satellites are designed to orbit between 500 and 2,000 kilometers above the Earth's surface, close enough for them to take high-quality images. The low orbit also makes them easier and cheaper to rocket into place.

There's a tradeoff, though: Unlike so-called "geosynchronous" satellites, which orbit from even higher up and remain in a fixed spot above the Earth, LEO satellites are constantly in motion, flying around the globe at around 8 kilometers per second – or 28,000 kilometers per hour.

Satellites in a time crunch

That means that as LEO satellites fly around the Earth they have only short windows, a few times per day, to beam the images they've collected to fixed antennae on the ground. "They gather lots of data, but don't have the time to transmit it," Kondrateva says. In a typical 15-minute transmission window, a satellite might be able to pass about only a few hundred megabytes of data down to Earth, depending on the exact configuration. The rest – possibly outdated by the next transmission window – is typically abandoned.

Kondrateva is working on ways to make the process more efficient. Right now, satellites leave the decisions about what images are important to analysts on the ground. By programming the satellites to pick and choose what data is worth transmitting, she hopes to help researchers make the most of each transmission window. "If you can't get all the data to Earth, you need to decide somehow what you want to transmit," she says. "I'm creating algorithms to let the satellite decide."

As part of her doctoral dissertation at HEIBRiDS, Kondrateva is in touch with engineers at the German Space Agency (DLR) and its European partners using LEO satellites to help people on Earth. The approach is typical at HEIBRiDS, which enrolled its first group of 13 Doctoral researchers in 2018. HEIBRiDS works with DLR and five other partners to offer Doctoral researchers input from different institutions and disciplines. The school is part of the Helmholtz Information & Data Science Academy (HIDA), Germany's largest post-graduate education network for data and computer science.

Disaster response thanks to satellite imagery

Working with partners at DLR, Kondrateva learned that satellite imagery is critical for disaster response: DLR satellites can spot and respond to wildfires, identify infrastructure damaged by earthquakes and tsunamis, and assess the damage from catastrophic floods.

When severe wildfires swept across California a few years ago, a DLR satellite mission called FireBIRD (Fire BiSpectral InfraRed Detector) monitored the temperature and intensity of the blazes and provided emergency responders in California with maps of the hottest fire areas. "I want to make all that more efficient," Kondrateva says. "It may take too long to send the images to the ground for analysis. The more information you can get on a satellite directly, the better."

The payoff could be huge – updated maps in hours, rather than days. "A lot of the data satellites collect is poor quality, or not needed," she says. "If you're doing fire detection, for example, you don't always need a picture of the fires, you just need to know where they are."

That's a problem because the models usually used to classify images, called neural networks, are both large and computationally demanding. On board a satellite, computing power is a precious commodity – and any new models also have to be transmitted up to the satellite, in the tight windows when it's in range. To cope, "I'm working on an algorithm to compress the models to make them smaller and need less computational resources," Kondrateva says.

A helpful linguistic background

At first glance, the leap from linguistics to low Earth orbit may seem huge. But Kondrateva says it was a natural progression: She learned computer programming as part of her linguistics research. She soon shifted her focus from human languages to computer languages, earning computer science degrees from Humboldt University and working as an engineer. “In the beginning I was trying to understand how languages worked internally,” she explains with a smile. “I was most interested in systems.”

Part of Kondrateva’s master’s degree dealt with how networks of satellites communicated with each other. When it came time to choose a PhD topic, she gravitated to space applications once again – and to HEIBRiDS, which gave her an opportunity to apply her computer science background in a new way. “I found data science exciting and wanted to learn something new,” she says. “Computer science, data science and satellites are a very nice combination.”

Kondrateva’s wide-ranging background makes her a good fit at HEIBRiDS. The school’s interdisciplinary mix has been helpful as the St. Petersburg native moves into yet another new area of research.

Like her fellow Doctoral researchers, Kondrateva has advisers from two different institutions to ensure interdisciplinarity. HEIBRiDS students meet every two weeks for PhD seminars and data science lectures, and follow a common curriculum as they work towards their doctorates.

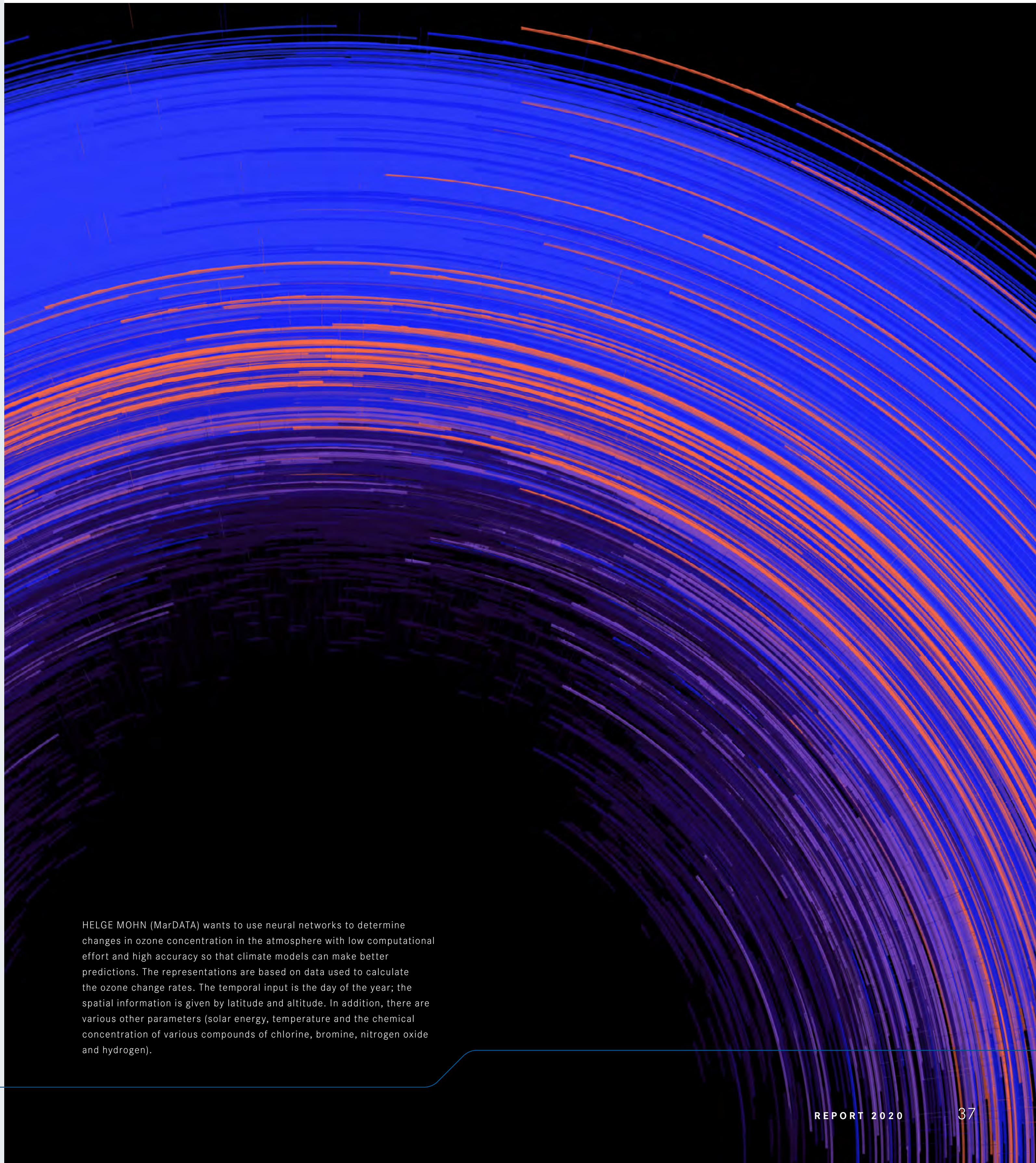
Kondrateva says colleagues from other disciplines help her pinpoint key data science research questions and avoid duplicating work already done by other scholars.” At HEIBRiDS if anybody has a problem, they can always ask,” she says. “The communication is really good.”

Text: Andrew Curry

https://www.helmholtz-hida.de/en/activities/news/newsdetail/helping-satellites-connect/

Credits: Olga Kondratevar: private/Firebird-Satellite: DLR

MarDATA HELMHOLTZ SCHOOL FOR MARINE DATA SCIENCE



HELGE MOHN (MarDATA) wants to use neural networks to determine changes in ozone concentration in the atmosphere with low computational effort and high accuracy so that climate models can make better predictions. The representations are based on data used to calculate the ozone change rates. The temporal input is the day of the year; the spatial information is given by latitude and altitude. In addition, there are various other parameters (solar energy, temperature and the chemical concentration of various compounds of chlorine, bromine, nitrogen oxide and hydrogen).

HELMHOLTZ SCHOOL FOR MARINE DATA SCIENCE (MARDATA)

The Helmholtz School for Marine Data Science (MarDATA) pools scientific marine expertise in the far north and offers young scientists the unique opportunity to refine Data Science methods specifically for the marine research. MarDATA has the goal of defining a new type of "Marine Data Scientists" and training them in a structured doctoral program. Scientists from computer science, information technology and mathematics work together on marine topics. This includes modelling on supercomputers, (bio-)computer science and robotics or statistics and big data methods.

Research areas:
The doctoral researchers at MarDATA conduct research in information technology, computer science or mathematics, each in connection with current issues in marine research.

Partners of MarDATA
GEOMAR Kiel, Alfred Wegener Institute Bremerhaven, Christian-Albrechts-University of Kiel, University of Bremen, Jacobs University Bremen.

Universität Bremen

JACOBS UNIVERSITY

GEOMAR

AWI

C | A | U Christian-Albrechts-Universität zu Kiel

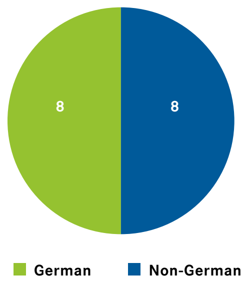


Fig. 24: Recruitment of doctoral researchers from abroad at MarDATA 2019-2020.

Applicant situation and recruitment

MarData currently has 16 PhD candidates. Half of them are women, also half of the PhD candidates come from abroad. In the first round of applications in 2019, 87 applications were submitted, 32 from Germany, the rest from international applicants. Eight women are represented among the PhD candidates.

The second round of calls for applications closed on February 17, 2021.

Events and networking

In addition to a Kick-off event at GEOMAR Kiel in 2019, a total of 32 events have taken place at MarDATA in the following year 2020:

5 courses

- as part of the advanced Scientific Programming Block Course
- Insights into Python 1, 2 & 3 (46, 41 and 39 participants)
- Software Testing Practices (26 participants)
- Data Science with Dask (33 participants)

14 lectures

- MarDATA Online Lecture Series (MOLS): Intro to Physical Oceanography
- MOLS: Intro to Climate Science
- MOLS: Intro to Marine Chemistry
- Intro to Marine Geology
- Intro Marine Biology
- Intro to Bioinformatics
- Intro to Marine Geophysics
- MOLS: Research Software Engineering

(all part of the Block Course "Introduction to Marine Science")

- Numerical Methods (52 participants)
- Advanced Statistics (54 participants)
- Software Licensing (33 participants)
- Introduction to Databases (30 participants)
- Software Testing Practices (37 participants)
- Modular Research Software (26 participants)

9 regular Digital Science Mondays (MarDATA and networking events)

- "MarDATA - Defining, Educating and Establishing a New Profile of Marine Data Scientists" (E.Prigge)
- "ESM Virtual Field Campaigns" (by W. Rath, MarDATA Supervisor)
- "How Does It Fit? a Quickview on Digitalization, Datahub, and Incubator Activities " (by S. Frickenhaus, MarDATA SB)
- " Data Mining in Very Large, Heterogeneous, Dynamic and Potentially Uncertain Data" (by M. Renz, MarDATA Supervisor)
- 5 MarDATA Project Intros by doctoral researchers

Other (MarDATA intern)

- Marine Data Science "Challenges & Opportunities" (25 participants)
- Interdiscplinary Supervison (23 participants)

Networking events

"MarDATA Weekly" is a weekly event that brings together all of MarDATA's doctoral researchers (in Bremerhaven, Bremen and Kiel); it was launched in March 2020 to maintain contact among PhD researchers during the Covid pandemic.

A special **networking event** was MarDATA's participation at the **HIDA Virtual Career Day** in 2020.

Publications

1. Razeghi, Y., Hasselbring, W., Berndt, C., & Dumke, I. (2020). Comparing Machine Learning Methods for Predicting Seismic P-wave Velocity on Global Scale. Proceedings of MACLEAN: MAChine Learning for EArth ObservatioN Workshop; paper 4. CEUR Workshop Proceedings, 2766.
2. Hiremath, D., Claus, M., Hasselbring, W., Rath, W. (2020) Automated Identification of Metamorphic Test Scenarios for an Ocean-modeling Application. 2020 IEEE International Conference On Artificial Intelligence Testing (AITest).
3. Mohn, H., Kreyling D., Wolthmann, I., Barschke, M., Rex, M. (2020) Estimating the Rate of Change of Stratospheric Ozone using Deep Neural Networks. AGU Fall Meeting 2020.

Further News from MarDATA

New cooperations are planned with the Data Science Center at the University of Bremen and with "Research Data Management and Data Science" from the University of Bremen Research Alliance.

Communication/Marketing

MarDATA has a **website** (www.mardata.de), a **Twitter account** with 203 followers (Feb 2021), and a **mailing list**.

SONAL RAMI – MARDATA

Modeling Ocean Currents

Even small ocean currents can have a significant impact on the climate. Sonal Rami's research at the MarDATA will help to predict these currents more accurately. By using machine learning methods, she is working on making complex ocean models even more precise - to enable more reliable predictions of climate change

"With every breath we take and every drop of water we drink, we are directly connected to the ocean," says data scientist Sonal Pravinbhai Rami. "Marine science connects directly to day-to-day life." It's that belief that brought Rami from a computer science teaching position at a university in India just a short drive from the warm waters of the Arabian Sea to the German port city of Bremerhaven in 2019.

For the computer science expert, Germany was a perfect spot to nourish two of her passions. "I'm a thalassophile. I'm fascinated by water," Rami says. "I spend all my time with computers and love to do coding in Python, but I'm also fond of the ocean."

That, Rami says, is part of the reason the Helmholtz School for Marine Data Science (MarDATA) seemed like a perfect fit when she decided to take her academic career to the next level. Now based in Bremerhaven, Rami is putting her skills as a teacher and machine learning expert to work improving sophisticated models of the world's oceans.

It's a critical job. Oceanographers have access to a steady stream of information, from sea surface temperatures and sea ice levels to ocean current speeds. With all that information flooding in, translating it all into meaningful predictions is a challenge. "In oceanography, there's more and more data coming in every day," Rami says. "With so much data, how can you gain knowledge from it?"

Optimize the ocean model FESOM

As part of her doctoral thesis, Rami is working on ways to interpolate sea surface temperatures to streamline oceanographic modeling. Her PhD focuses on improving the Alfred Wegener Institute's Finite-Element/volume Sea Ice-Ocean Model, known as FESOM. First introduced in 2004, it's one of the institute's most powerful climate modeling tools.

FESOM is essentially computer code capable of simulating how ocean currents all around the globe are impacted by sea ice. The way ocean currents interact is also part of the model – if water slows down in the Arctic, for example, there's a ripple effect that pulses through marine environments around the world.

Because more than 70 percent of the world's surface is covered by oceans, the behavior of their currents plays a huge role in the global climate. For oceanographers and climatologists, predicting climate change depends on understanding the oceans' contribution. Over the past decade, FESOM has contributed to climate change assessments put forward by the Intergovernmental Panel on Climate Change, considered the gold standard for climate change research.

It's no small job. Ocean currents are fiendishly complex. In the real world, ocean eddies – microcurrents that cumulatively add up to major trends – can be as small as 1 km and as wide as 25 km. But because of the sheer complexity of ocean systems, it takes lots of computing power to build accurate models. Computing limitations mean FESOM and other ocean models only calculate currents down to a resolution of 100 km, and often less.

By teaching computers to make predictions based on past models, Rami hopes to make it possible for FESOM to make accurate forecasts about the future using so-called neural networks. "My goal is to provide machine-learning solutions for oceanography and climate modeling," Rami says.

Interpolation can simplify mathematical modeling

She's focused on a technique called interpolation, which is like drawing a picture by connecting dots: Rather than draw all the lines each time the model is computed, Rami is teaching the computer to predict the movement of ocean currents faster and more efficiently based on a few data points. "Ideally," Rami says, "we don't need to store the middle data."

In other words, if sea surface temperatures in a particular region are 25 degrees in January and 27 degrees in March, "what about February?" Rami asks. "Using machine learning, we can more intelligently fill up the middle data" – and assume February's temperatures were about 26 for the purpose of the ocean model.

Of course, the real world is far more complex. But with enough data, "you plug in variables" – sea surface height, temperature, salinity, and wind velocity, for example – "and get predictions," Rami says. "It's a cheap but effective solution that uses data-driven approaches for the forecasting and reconstruction of sea surface dynamics."

Rami's PhD is supervised by researchers from the Alfred Wegener Institute, the GEOMAR Helmholtz Centre for Ocean Research in Kiel, and the University of Bremen. Such collaboration is an example of what MarDATA – which brings together scientists from the Helmholtz Association, the Alfred Wegener Institute Helmholtz Center for Polar and Marine Research and several German universities – does well.

With 15 doctoral researchers under the supervision of researchers from different institutions, the school is part of the Helmholtz Information & Data Science Academy (HIDA), Germany's largest post-graduate education network for data and computer science.

Research at MarDATA: "Great background"

The interdisciplinary opportunity to combine data science and oceanography wasn't the only reason Rami applied to MarDATA. As an assistant professor in India she worked on climate models for the Indian Space Agency, sparking an interest in the topic.

And her first experience in Germany, at a series of expert lectures on machine learning run by the Max Planck Institute for Intelligent Systems in Tuebingen, was overwhelmingly positive. "I really got inspired," she says. "I thought, why not Germany for a PhD? It's a world leader in terms of technology, and German scientists are moving really fast when it comes to machine learning."

So far, the collaborative nature of MarDATA suits the former teacher well. "We have discussions every day, I meet with my supervisor twice a week," Rami says. "I get so many suggestions and great feedback on my work." With researchers spending most of their time at home as a result of the COVID-19 crisis, Rami's fellow doctoral researchers have even set up virtual common rooms to socialize.

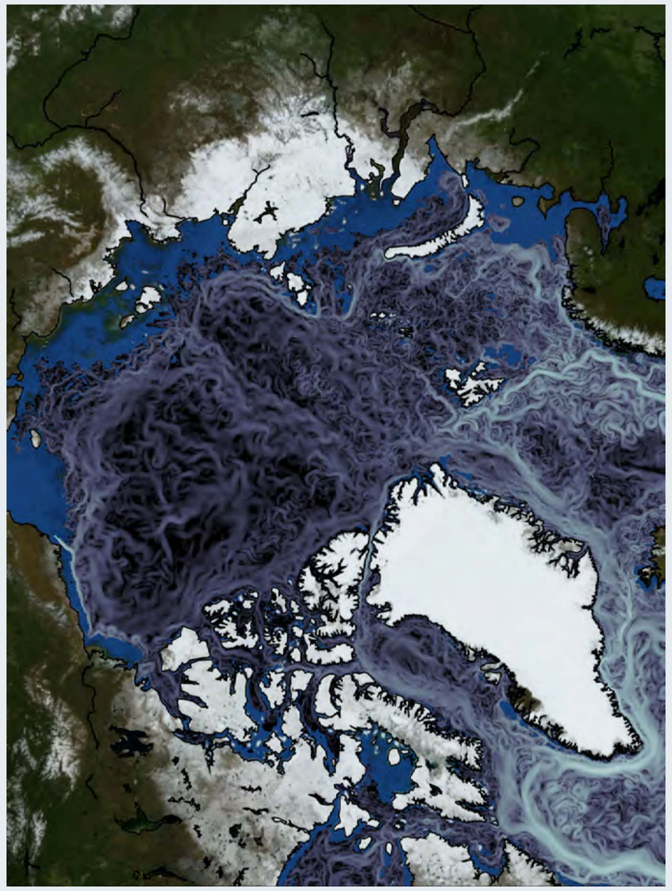
And although she arrived with a deep background in data and computer science, her time at MarDATA began with a block course that immersed doctoral researchers from different disciplines in the basics of oceanography and climate science. "We looked at ocean chemistry, biodiversity, physical oceanography, climatology – all great background," Rami says.

Her project is moving forward fast. She's already eyeing ways to interpolate variables beyond sea surface temperature, like wave height. And she's convinced machine learning has tremendous potential when it comes to climate research beyond the oceans. "The long-term vision is to go beyond interpolation and create solutions for bigger climate models," Rami says. "Hopefully at some point there won't be any need to run physical models again and again, no need to store all the middle data. Machine learning can improve the realism of climate models."

Author: Andrew Curry

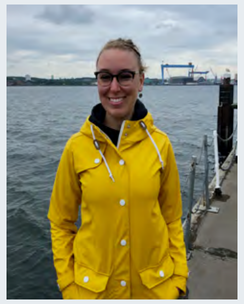
https://www.helmholtz-hida.de/en/activities/news/newsdetail/modeling-ocean-currents/

Credits: Sonal Rami: private/Visualization from the FESOM1.4 ocean model: Dmitry Sein (simulation), Nikolay Koldunov (visualization)



MARIA-THERESIA VERWEGA – MARDATA

Math for the Marine Environment



Maria-Theresia Verwega is a mathematician with all her heart and she loves the sea. Early on, the Kiel native felt the need to contribute to marine protection with her knowledge. That's why she decided to write her PhD at MarDATA - a new, interdisciplinary research school of the Helmholtz Association where marine data scientists are trained. Because they are urgently needed in ocean and climate research

The GEOMAR Helmholtz Centre for Ocean Research Kiel is located in Düsternbrooker Weg, directly next to Kieler Förde, where a few seagulls are always circling. Maria-Theresia Verwega is already standing outside the door, waiting for the visitor to whom she wants to show her new research field. She has had an office at GEOMAR for a few months now; she has a second office at the Christian-Albrechts-University in Kiel. The first thing that stands out about the young doctoral researcher? Her hooded sweater - with a seagull on it - and the glow on her face, which only people have who have found the right place for themselves. "To be here," she says and smiles, "feels like a dream come true. I always wanted to do oceanography, but I'm a mathematician."

Maria-Theresia Verwega, born in 1989, friends and colleagues call her „TiA", was born in Kiel - "in the university hospital, with a view of the sea. Since then the element has been part of her life. Tia doesn't like swimming in lakes "because the water doesn't taste salty." Every year from March to November she jumps into the Baltic Sea at Falckensteiner Strand - no matter how cold or windy it is. Recently, she even goes swimming all year round. "The sea," she says and sounds almost like she's in love, "is simply always beautiful."

She discovered her talent for mathematics at an early age

But the sea does not remain her only passion: she discovered her talent for mathematics at an early age. She loves the clarity of it and the tinkering until the right solution is found. That's why she decides to study mathematics at the Christian-Albrechts-University of Kiel after school. Minor subject: physical geography. "Because I always had the urge to do something tangible with my knowledge," she says, "something useful for the world."

After her bachelor's degree, she takes two masters on top: one in mathematics and one in environmental management. On the mathematical side, she specializes in modeling, optimization, numerics, and in environmental management in ocean research. Her master's theses in both subjects are on mathematical methods of data analysis and biogeochemical modelling.

In simple terms, biogeochemical modelling is the representation of the interrelationships of chemical, biological or physical processes in order to draw conclusions from them. Alfred J. Lotka's and Vito Volterra's predator-prey models, for example, are well known. From the observation of different animal populations or plant occurrences that occur together, the researchers derived rules on how these interact.

Understanding the processes in the ocean

In ocean research, these interactions are considered at the smallest level: Using so-called NPZD models, the dependencies between nutrients, phytoplankton, zooplankton and detritus are observed in order to draw conclusions about ocean processes from these cycles and to use these data to calculate the future - for example, to make climate predictions. "And that's exactly what pure mathematics is," says Tia. "Because there are differential equations in the computer programs."

Finally she could apply her math skills to the sea. But the symbiosis of her passions in her life did not stop there: She had just handed in her master's thesis at the beginning of 2019 when she heard about the idea of a "Helmholtz School for Marine Data Sciences" - in short: "MarDATA". An interdisciplinary research school that was dedicated to training a new type of marine data scientist. A path that she had already begun to follow.

She strongly hoped for the approval of the research funds - and applied as soon as MarDATA was launched. The Research School, which has been in existence since autumn 2019, is a joint foundation of the two leading German marine research institutions - the Alfred Wegener Institute (AWI) in Bremerhaven and GEOMAR in Kiel - and their partner universities in Bremen and Kiel. It is also part of the Helmholtz Information and Data Science Academy (HIDA), Germany's largest postgraduate education network in the information and data sciences. MarDATA's doctoral researchers, who come from the fields of computer science, mathematics and engineering, receive three-year contracts and are supervised by two professors one of each from the data and marine sciences.

Countless measuring instruments collect data on the world's oceans every day

Tia was granted a position. She has been part of the first cohort of 16 doctoral researchers since September 1, 2019, who are researching at AWI or GEOMAR and who are learning to integrate their data expertise even better with their knowledge of marine processes. Because this is exactly what is urgently needed today: Over 70 percent of the Earth's surface is covered by oceans. Countless measuring instruments have long been in daily use on all oceans, whether on expedition ships or in the form of autonomous mini-submarines that permanently generate data on water temperatures, salinity, nutrient concentrations and much more. In short: a huge treasure trove of data, which contains enormous knowledge potential - as long as you know how to handle it.

"Our task at MarDATA is to process this data in a useful way," says Tia. In her doctoral thesis with the title: "Development of adapted Kernel Density Estimators for the Calibration of marine biogeochemical models", she is now exploring how to improve predictions for Earth system models. In other words, the graphs that we may know from the news or the IPCC report, which give predictions of how our climate or certain processes in the world will change in the future and under what conditions. The basis for decision makers such as politicians or managers.

"I'm not a Greta and I can't lead masses. I am a mathematician. Now I can use it for what I always wanted to do."

Tia optimizes a biogeochemical submodel for this purpose. She analyses datasets using a core density estimator, creates metrics and recalculates the model using different parameters until she finds solutions for better predictions. "I'm not a Greta and I can't lead masses, and I'm not a marine biologist and I don't go out into the Antarctic in storms and wind," Tia - who is not entirely seaworthy by the way - says. "I am a mathematician. I'm really good at it. And now I can finally use it for what I always wanted to do." For protecting the sea that she loves.

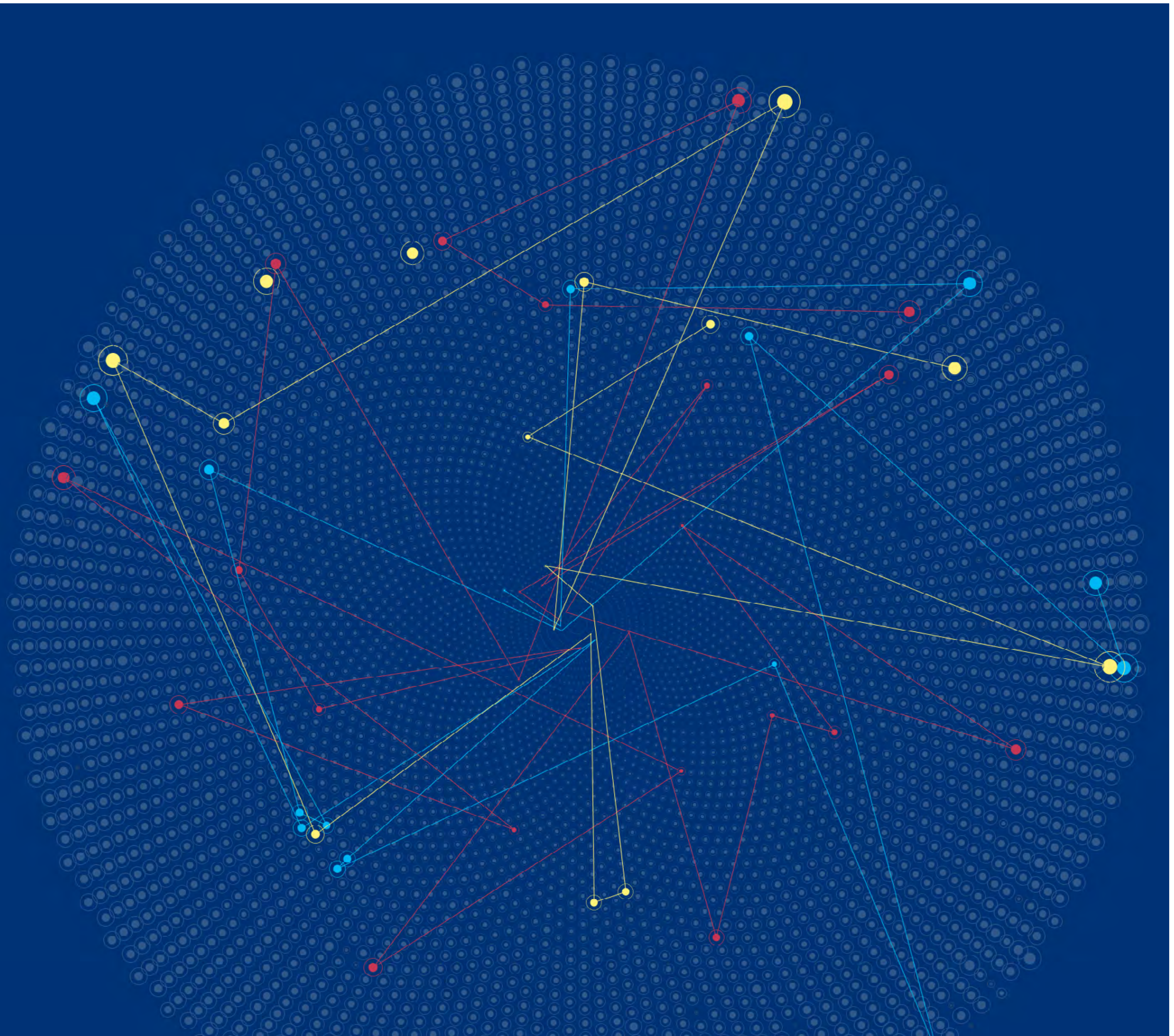
Author: Andrea Walter

https://www.helmholtz-hida.de/en/activities/news/newsdetail/math-for-the-marine-environment/

Credits: Maria-Theresa Verwega: private

DASHH Data Science in Hamburg
HELMHOLTZ Graduate School
for the Structure of Matter

DATA SCIENCE IN HAMBURG – HELMHOLTZ GRADUATE SCHOOL FOR THE STRUCTURE OF MATTER (DASHH)



STASIS CHUCHURKA (DASHH) uses data science methods to conduct research into the development of a single model for analyzing X-ray data in order to gain a better understanding of the electronic structures of substances and their subsequent chemical processes.

The data used for this picture pertain to the electronic charging density of urea. They comprise the density matrix elements and two corresponding atomic orbitals each.

The DASHH Data Science in Hamburg - Helmholtz Graduate School for the Structure of Matter is an example of the excellent work of the Helmholtz Schools. In the School, DESY cooperates with numerous institutions from Hamburg and other northern German states to provide innovative training and cooperation for doctoral researchers in the field of Data Science. With its approach, DASHH is pioneering the development of new collaborative approaches to evaluate complex, heterogeneous data using intelligent algorithms. In 2020, DASHH was awarded the North German Science Award for its excellent work and pioneering qualities in connecting universities and research institutions.

Research areas:
Doctoral researchers at DASHH deal with questions from a wide range of research fields, such as structural biology, particle physics, materials science and science with ultra-short X-ray pulses.

Partners of DASHH
Deutsches Elektronen- Synchrotron DESY, Hamburg University of Technology, University of Hamburg, Helmut Schmidt University Hamburg, Helmholtz Zentrum Geesthacht, Helmholtz Center for Infectious Diseases, Max Planck Institute for Structure and Dynamics of Matter, European XFEL

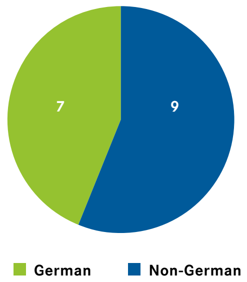


Fig. 19: Recruitment of doctoral researchers from abroad at DASHH 2019-2020, (excl. associated doctoral researchers).

Applicant situation and recruitment

Eight new PhD positions were awarded in each of 2019 and 2020: 208 people interested in the program applied in 2019; in 2020, 118 candidates submitted applications. Applicants came from all over the world: in 2019, interested candidates from 45 countries applied; in 2020, applicants came from 27 countries.

A total of 16 doctoral researchers from ten countries are currently enrolled at DASHH with a dissertation project, of which three are women and 13 are men. In addition, there are two male associated doctoral researchers from DESY.

Events and networking

There was a significant increase in this area in 2020: while there was a total of six events in 2019 (four networking events and two lectures), 29 events were carried out in 2020 (11 networking events and 18 lectures). DASHH initiated a Hamburg COVID-19 Lecture Series in collaboration with the Leibniz Science Campus InterACt and also initiated a monthly Data Science Colloquium. Monthly Get Together events with all DASHH doctoral researchers and associated doctoral researchers were established at the end of 2020.

8 PhD Seminars

- 24.01.2020, Prof. Schwanenberger, The Dark Side of the Top
- 24.01.2020, Jonas Rübenach, The Search for Exotic Heavy Higgs Bosons with the CMS Experiment at the LHC
  08.05.2020, Gianluca Martino, Steps towards a Self-Healing LLRF System
- 08.05.2020, Dr. Schlarb, Intelligent Process Control for Accelerators at the Example RF Controls
- 29.05.2020, Prof. Rarey, Computational Molecular Design - Models, Algorithms, Software
- 29.05.2020, Theresa Cavasin, Outlier Detection for Improved 3D Reconstruction in Single Particle Cryo-EM
- 19.06.2020, Dr. Meents, X-ray Protein Crystallography under Non-ambient Conditions
- 19.06.2020, Michael Größler, Serial Crystallography: Using Bad Crystals for Better Resolution
- 03.07.2020, Georgiana Mania, A Framework for Efficient Parallelized Execution of Charged Particle Tracking at the Luminosity Frontier
- 03.07.2020, Dr. Styles, An Introduction to High Energy Physics Track Reconstruction
- 08.11.2020, Prof. Küpper, Unraveling Molecules at Work
- 08.11.2020, Yahya Saleh, Active Learning of Molecular Potential Energy Surfaces
- 27.11.2020, Tom Weber, Software Engineering Models for Error Mitigation in Quantum Computing - Part II
- 27.11.2020, Prof. Riebisch, Software Engineering Models for Error Mitigation in Quantum Computing - Part I
- 11.12.2020, Prof. Bause, Space-Time Finite Element Approximation of Waves
- 11.12.2020, Nils Margenberg, THz Generation in Periodically Poled Nonlinear Crystals

9 Hamburg COVID-19 Series

- 12.08.2020, Prof. Addo, COVID-19 Vaccine Development: Where Do We Stand?
- 26.08.2020, Dr. Lütgehetmann, COVID-19: Diagnostics and Pathological Findings
- 09.09.2020, Dr. Meents, Structure based Drug Design: Massive X-ray Screening of a Repurposing Drug Library against the SARS-CoV-2 Main Protease
- 23.09.2020, Prof. Baumbach, From Systems Medicine to COVID-19 - Network-enhanced Drug Repurposing
- 07.10.2020, Prof. Grünewald, The Coronavirus Guide to the Cell: A Structural Cell Biology Perspective
- 21.10.2020, Prof. Grundhoff, Genomics and Transmission Tracing of SARS-CoV-2
- 04.11.2020, Prof. Fangohr, The Open Science COVID Analysis (OSCOVIDA) Project
- 18.11.2020, Prof. May, The SARS-CoV-2 Epidemic Miracle in Sub-Saharan Africa
- 02.12.2020, Prof. Mnich, Algorithms, Network Epidemiology, Contact Tracing

Data Science Colloquium

- 03.12.2020, Prof. Simon, Big Data, Artificial Intelligence & Ethics

Networking Events

Two PhD Get Together meetings were organized in 2020 and are monthly reoccurring events. An additional match making event for DASHH PIs war organized in June 2020.

Publications

First author publications (journal/conference):

1. Martino, G., Riener, H., and Fey, G. (2020). Revisiting Explicit Enumeration for Exact Synthesis. Paper presented at: 2020 23rd Euromicro Conference on Digital System Design (DSD).

Co-authored publications (journal/conference):

1. Duwe, K., Lüttgau, J., Mania, G., Squar, J., Fuchs, A., Kuhn, M., Betke, E., and Ludwig, T. (2020). State of the Art and Future Trends in Data Reduction for High-Performance Computing. Supercomputing Frontiers and Innovations 7, 4-36.
2. Günther, S., Reinke, P.Y.A., Fernández-García, Y., Lieske, J., Lane, T.J., Ginn, H.M., Koua, F.H.M., Ehrt, C., Ewert, W., Oberthuer, D., et al. (2020). Inhibition of SARS-CoV-2 Main Protease by Allosteric Drug-binding. bioRxiv.
3. Günther, S., Reinke, P.Y.A., Oberthuer, D., Yefanov, O., Ginn, H., Meier, S., Lane, T.J., Lorenzen, K., Gelisio, L., Brehm, W., et al. (2020). Catalytic Cleavage of HEAT and Subsequent Covalent Binding of the Tetralone Moiety by the SARS-CoV-2 Main Protease. bioRxiv.
4. LaForge, A.C., Benediktovitch, A., Sukharnikov, V., Krušič, Š., Žitnik, M., Debatin, M., Falcone, R.W., Asmussen, J.D., Mudrich, M., Michiels, R., et al. (2020). Time-resolved Quantum Beats in the Fluorescence of Helium Resonantly Excited by XUV Radiation. Journal of Physics B: Atomic, Molecular and Optical Physics 53.

Further news from DASHH

DASHH participated in the Helmholtz Virtual Data Science Career Day (23.9.2020) with two PIs and two DASHH fellows presenting their research in virtual Zoom sessions and the coordinators available for discussion in Zoom during the whole day.

Conference Contributions:

- 29.-31.01.2020, 2020 European XFEL Users' Meeting, Poster Contribution
- 23.-25.09.2020, FSP CMS Workshop, Talk Contribution
- A DASHH fellow gave a talk at the ACTS Tracking for High Energy Physics Workshop on 25.05.2020.

Awards:

- DASHH is the laureate of the North German Science Award 2020, which is awarded every two years by the North German Conference of Science Ministers to promote special collaborative projects. It is endowed with 125,000 euros.
- Jonas Rübenach won the CMS Deutschland Photo Contest 2020 (1st place). It was announced during the FSP CMS Workshop 2020, September 23-25, 2020 (online).

New Collaboration:

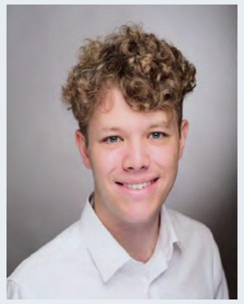
- The Hamburg University of Applied Sciences (HAW) was gained as a new partner institution; in addition, collaborations were established with the Center for Data and Computing in Natural Sciences (CDCS) at the University of Hamburg and with the PIER Education Platform (PEP) for career, research and soft skills.

Communication/Marketing

DASHH has a **website** (www.dashh.org) and maintains a **LinkedIn account** with 204 followers. Recent news is distributed via the corresponding **mailing lists** as well as through the partner organizations and collaboration partners (the Center for Data and Computing in Natural Sciences (CDCS), the clusters of excellence CUI - Advanced Imaging of Matter and Quantum Universe of the University of Hamburg (UHH), the Department of Informatics of the UHH, the Institute of Mathematics of the Hamburg University of Technology and ahoi.digital - the Alliance of Hamburg Universities for Computer Science). DASHH also advertises the regular Calls for Application on various websites, e.g., DAAD, LinkedIn, nature jobs, pro-physik, Physics Today, Science Careers, academics.de, academics.com, jobs.zeit.de

TOM WEBER – DASHH

Basketball Player at the Super Computer



Quantum computers promise unimaginable processing power — but only in theory. So far, they can't be used because their results are rife with errors. In his doctoral thesis, Tom Weber is working at DASHH on a software solution that aims to compensate for these errors

And now he has a dog to take care of, too. "His name is Marley, he's a poodle - West Highland terrier cross," says Weber, pointing to the white creature lying next to him in its dog basket. His girlfriend's parents gave them Marley to take care of, so at least Weber gets to take long walks, even if his basketball training has been canceled for months due to the coronavirus pandemic.

Weber is easy to talk to. This first thing you notice about him is his distinctive curly hair, which makes him appear taller than he already is with his classic basketball player's stature of over six feet. But most of all, you notice the ease with which he suddenly switches from talking about his dog or sports to the highly complex topic he's most preoccupied with right now — quantum computers. "I had absolutely no plans to work on this topic," he says, "but the first time I read about it, I was immediately fascinated." The official title of his doctoral thesis is "Optimization of quantum circuits using error analysis and modeling" — but Weber quickly follows this up with an explanation that sheds a bit more light on the topic. "Right now, quantum computers have high error rates. In my project, I'm looking for methods that will make it possible to use them in the near future despite these errors." The software Weber is developing can be imagined as a type of filter: It filters the results quantum computers produce during their calculations and removes the errors from them.

Faulty qubits

But Weber's penchant for complex topics started before his doctorate. He grew up between the cities of Hamburg and Lüneburg in northern Germany and was still in school when he realized that he was fascinated by the questions covered in his physics classes. Doing a degree in physics was something he had never considered. During his master's in Hamburg, he focused on mathematical aspects in the field of physics. "I had actually planned to do a doctorate in mathematics — but I applied for many, many doctoral positions, and nothing came of it." But then he happened to come across a tender for this project at DASHH (Data Science in Hamburg – Helmholtz Graduate School for the Structure of Matter), one of six Helmholtz Information & Data Science Schools. All of the schools are part of the Helmholtz Information & Data Science Academy (HIDA), Germany's largest post-graduate training network in the field of information and data science. The tender was for a doctoral researcher who would be working with quantum computers. "I read the outline of what it involved and liked it so much that I applied there and then." Weber got the position and started his work when he was just 24 years old — he had skipped first grade when he was a child, which meant he was only 17 when he began his degree.

And how many times already has he explained exactly what he's researching? "Countless times," says Weber without thinking twice. When he explains his project, he usually starts with how a quantum computer works. While a conventional computer processes bits consisting of zeros and ones, quantum computers use quantum bits, known as qubits. Qubits can not only be zeros and ones but also any imaginable value in between. This means they can process more computing operations simultaneously — and experts believe that they will soon eclipse even today's supercomputers, which work with conventional bits. "The only problem is that qubits still throw up a lot of mistakes," says Weber. "If you think about it in terms of conventional computers, it would be as if it suddenly displayed a one instead of a zero." This is the starting point for his research project, and the goal is to eliminate these errors from the results.

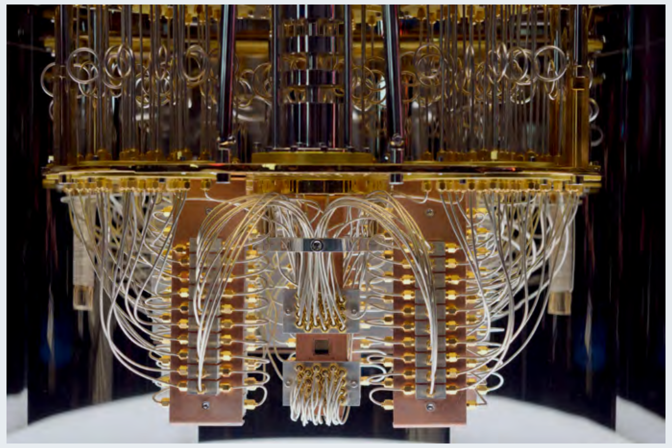
He has never stood before a quantum computer

It's also important to understand that science is pinning a lot of hopes on quantum computers right now. Behind the scenes, a global competition is underway, and big companies are also involved: IBM, for instance, is conducting intensive research on quantum computers, Google is pursuing an extensive research program, and Chinese firms such as Alibaba and Baidu are also registering an ever-growing number of patents. At the heart of this competition is the question of how many qubits can already be built into a computer today — that is, how powerful the computers are. Remarkable advances are attained on a regular basis, but the problem is still the error rate.

"Science is approaching this challenge from two different angles," Weber explains: "Some researchers focus on reducing the error rate. The main aspect for them is improving the hardware." And then there's the other angle — which Weber works on in his own project. "We accept that quantum computers aren't all that reliable yet and try to make the faulty results they produce usable despite that." So far, Weber has never stood in front of a quantum computer himself. When he uses their processing power, he does so from home or from his office — which involves sending tasks to the quantum computers (he currently uses devices at IBM in the state of New York in the US) and receiving the results a short time later.

A software to check the super computers

Once again, Weber uses a tangible comparison to explain his work: "Imagine a dice," he says. "If you roll the dice once and you get a four, that's a snapshot. A picture only emerges when you roll the dice many times in a row and take a look at the results. You can see whether there are deviations and anomalies." In terms of Weber's work, this means that quantum computers run through the same task many times. And Weber is developing a piece of software that runs on a conventional computer and analyzes the quantum computer's results to check for these deviations and anomalies. "I do this by having the quantum computer process a task for which I know the result. This lets me verify whether my program is drawing the right conclusions based on the results from the quantum computer."



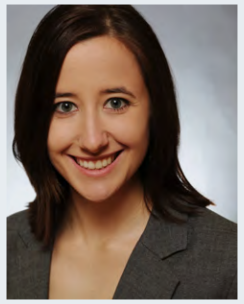
It's a highly complex mathematical problem — and this is where Weber's original plan of doing a doctorate in mathematics comes full circle. He still remembers well the crucial conversation before he got the position at the graduate school. "I was sitting across from a math and physics professor, a doctoral researcher from the team I work on now, and two postdocs, and they all knew that I hadn't had anything to do with quantum computers before that." Nonetheless, he got the position thanks to his mathematical skills — and because the chemistry was right. Which is another important aspect, because after all, Weber doesn't do his research in isolation. He is involved in various working groups throughout the week, both at the University of Hamburg and at DESY in Zeuthen. "In all these meetings, we look at individual aspects that play a role in my work," says Weber, "and I come away with a lot of knowledge wherever I go."

So much knowledge that he sometimes needs to work off some steam. And just like that, the conversation is back to basketball and Marley the mutt. Weber often plays piano too when he's done with work." I started playing when I was seven years old," he says. Right now, he says he's working his way through Frédéric Chopin's Fantaisie-Impromptu — and, unlike basketball, he hasn't had to take a break from the piano due to coronavirus.

Author: Kilian Kirchgeßner

https://www.helmholtz-hida.de/en/activities/news/newsdetail/basketballer-am-superrechner/

Credits: Tom Weber: private/IBM quantum computer: Boykov/shuterstock.com



ANNA THERESA CAVASIN – DASHH

Better Medicines Thanks to IT

For the development of new medical agents, it is important to know how biomolecules, for example viruses, are structured. Bioinformatician Anna Theresa Cavasin therefore wants to use computer-assisted methods to optimize images from cryoelectron microscopy - in order to pave the way for better drugs with data science

The bicycle racks in front of the Centre for Structural Systems Biology, or CSSB for short, stand empty on this sunny fall morning. Offices and labs at the interdisciplinary infection research center in the Bahrenfeld quarter of Hamburg are sparsely occupied due to the coronavirus pandemic. Anna Theresa Cavasin, a bioinformatician and doctoral researcher, is wearing a face mask when she comes down to the building's revolving door; she then leads us up an elegantly curved staircase made of light wood to the conference area on the second floor. We meet a few colleagues on their way to the coffee machine, but otherwise, everything is quiet in the bright new building that opened in 2017 here on the German Electron Synchrotron (DESY) campus.

The CSSB actually has enough space for 180 scientific staff. In more normal times, Anna Theresa Cavasin works three days a week here together with other researchers in one of the open-plan offices connected to the lab. The 26-year-old is wearing black jeans and a checkered flannel shirt; her gaze is focused and alert. For nine months now, Cavasin has been one of 14 doctoral researchers at the DASHH graduate school, which is officially known as Data Science in Hamburg – Helmholtz Graduate School for the Structure of Matter. The school – part of the Helmholtz Information & Data Science Academy (HIDA), Germany's largest postgraduate training network in the field of information and data science – is a hub for projects led by young scientists who conduct basic research into the structure of matter. Cavasin's project is called "Next generation integrative modeling for cryoelectron microscopy."

The goal: to improve the microscopic image quality of biomolecules

While its title might sound somewhat abstract, the project pursues a very tangible goal: Cavasin wants to improve imaging results in cryoelectron microscopy, an increasingly relevant method of determining the structure of biomolecules. In simple and specific terms, she does this by applying an algorithm to automatically sort out poor-quality results with little informative value from the vast quantities of images that are produced when working with a microscope. In the over nine months since she became a doctoral researcher at DASHH, the young bioinformatician has been working to select parameters that differentiate good images from those of poor quality. Once she has determined these variables, Cavasin will spend the remaining part of the three years planned for her doctorate working to program algorithms that sort image material accordingly.

This process is somewhat more complex in scientific terms, of course. So, it's worth taking a look at the starting point of Cavasin's research. Understanding the structure of certain biomolecules to the most accurate extent possible is a basic area of knowledge that is required in the development of drugs in particular. Cryoelectron microscopy is increasingly being used to arrive at valid data relating to these smallest of particles. Cavasin works with frozen rather than crystallized samples – which, among other things, has the advantage of enabling the analysis of proteins that cannot be crystallized. The resolution of the images produced by cryoelectron microscopes has increased significantly over the past ten years – in fact, those responsible for developing the method were awarded the Nobel Prize for Chemistry for their contributions in 2017. But, Cavasin says, there are still many cases in which the quality of the images is insufficient.

Why would scientists need hundreds of thousands to millions of 2D microscopic images of a single molecule? Because they are the basis for the incredibly large-scale process of calculating 3D models – for instance, of a certain protein. One thing is for certain: The better the individual 2D images, the more exact the 3D image produced from them. And what happens next with the 3D model? Cavasin explains: "We want to see what a protein involved in an infection looks like, down to the level of individual atoms if possible. This is the only way we can develop active ingredient molecules – in other words, drugs – against a virus, for example."

"I think informatics is the best way to approach questions in biology"

Even when she was in school, Cavasin was very interested in biology and chemistry, and especially in chemical and molecular biology. She entered school competitions together with two of her friends, who are now working toward their doctorates in the sciences as well. On the website of her school in Essen, where she grew up, there are still images of her smiling into the camera as she holds a certificate. She carried out various analyses back then, including for her presentation at the International Biology Olympiad, "Onion Cells in the Ion Trap" – which earned her a place among the top twelve participants from North Rhine-Westphalia. Did her teachers notice and encourage her talent? Not really, Cavasin says; she did a lot of reading herself, in books and online. But what really motivated her, she says, were her friends. They were also the ones who drew her attention to the opportunity of starting university early when she was 17. And this was how Cavasin started regularly attending university lectures in her last year of high school rather than English classes, which she describes as boring.

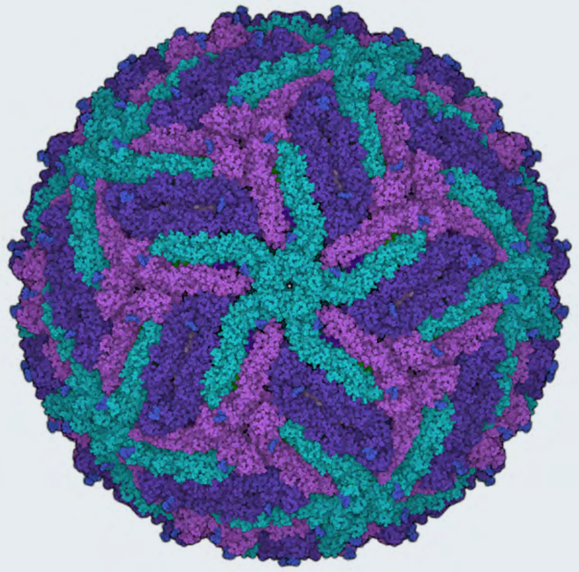
After completing high school, she then began studying toward a bachelor's degree in chemical biology in Dortmund. Given where she is today, it's clear that her thesis on computer-assisted active ingredient design determined her next move, which took her to Hamburg for a master's degree in bioinformatics. Cavasin studied under Matthias Rarey, Head of the Center for Bioinformatics at the University of Hamburg – and one of the spokespersons for the new DASHH graduate school. Rarey is now supervising Cavasin's doctoral thesis as well. In addition to her desk at CSSB on the DESY Campus, she also has a workstation at the university. She says that working in bioinformatics lets her get to the heart of what she's really interested in: "I think informatics is the best way to approach questions in biology, because it's so structured. I like structures. And I like efficiency." A sense of curiosity is something that her group at the graduate school shares as well, she says, noting that their main question is always: What's the reason for that? "Gaining new knowledge is our common goal."

Cavasin says that she sees the world in a rational way. She wants to understand – down to the smallest detail, wherever possible. "I've always been interested in what's behind processes. Why the world works the way it does." She's happy that she has ended up doing her doctorate in the special field of cryoelectron microscopy. "The job we do here isn't one that a lot of people do – and I think that understanding something like this at a very fundamental level at some point is amazing. But only as long as I feel that what I'm doing is worthwhile – and that my work ultimately helps people in a tangible way." Such as giving them access to better drugs, for example.

Author: Christiane Langrock-Kögel

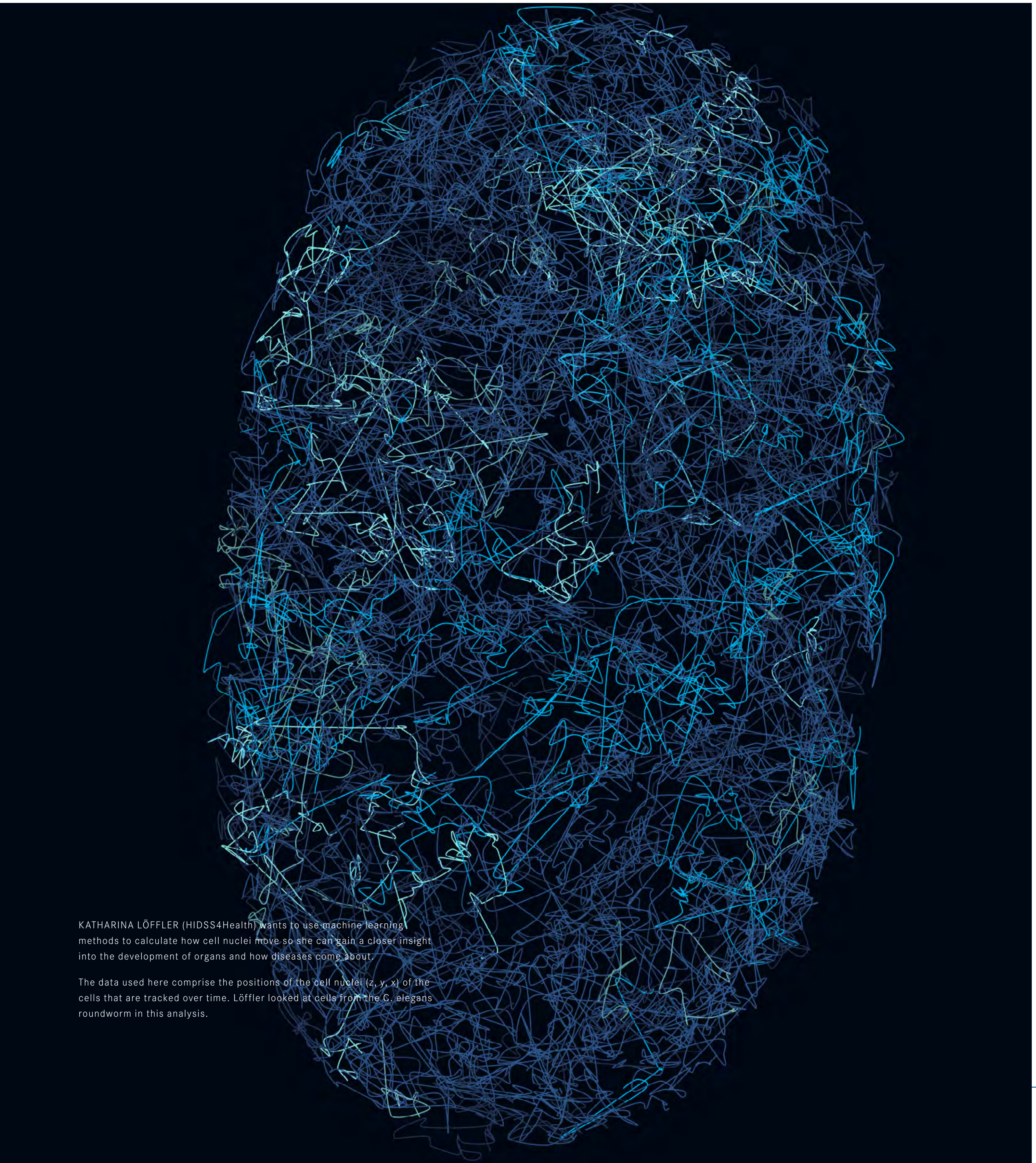
https://www.helmholtz-hida.de/en/activities/news/newsdetail/better-medicines-thanks-to-it/

Credits: Anna Theresa Cavasin: private/ Representation of a Zika virus based on a cryoelectron micrograph: StudioMolekuul/Shutterstock.com



HIDSS 4HEALTH | HELMHOLTZ INFORMATION & DATA SCIENCE SCHOOL FOR HEALTH

HELMHOLTZ INFORMATION & DATA SCIENCE SCHOOL FOR HEALTH (HIDSS4HEALTH)



KATHARINA LÖFFLER (HIDSS4Health) wants to use machine learning methods to calculate how cell nuclei move so she can gain a closer insight into the development of organs and how diseases come about.

The data used here comprise the positions of the cell nuclei (z, y, x) of the cells that are tracked over time. Löffler looked at cells from the C. elegans roundworm in this analysis.

The goal of HIDSS4Health is to train and promote the best young talent at the interface between Data Science and health. HIDSS4Health offers a structured doctoral program with close integration into an interdisciplinary network of leading scientists from both the data and life sciences.

Research areas:
Imaging & diagnostics, surgery & intervention 4.0 and models for personalized medicine.

Partners of HIDSS4Health:
Karlsruhe Institute of Technology, German Cancer Research Center Heidelberg and University of Heidelberg

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

KIT Karlsruher Institut für Technologie

dkfz. DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

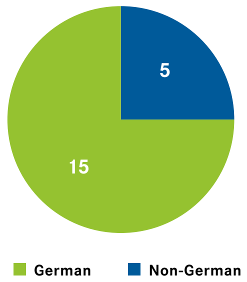


Fig. 22: Recruitment of doctoral researchers from abroad at HIDSS4Health 2019–2020 (excl. associated doctoral researchers).

Applicant situation and recruitment

There are currently 20 PhD positions filled at HIDSS4Health, 11 of which were awarded in 2019 and the other nine in 2020. In 2019, there were a total of 223 applications, of which 173 were not from Germany. In 2020, 147 applications were received, 108 of which came from abroad.

Of the 20 doctoral researchers, 75% (15) are German, and the proportion of women is 45% (9). Three additional female researchers and two additional male researchers are associated with HIDSS4Health; they are from KIT (1), DKFZ (3), and Heidelberg University (1).

Events and networking

In 2019, HIDSS4Health hosted 13 events (3 networking, 10 lectures); in 2020, it organized 28 events (12 networking events, 1 course, 13 lectures, 2 other):

12 Networking events

- 08.04.2020, Selection Event 2020, Online (41 participants, 17 HIDSS4Health, 5 Helmholtz, 19 external, 16 female): Selection Event with talks from the applicants, short project pitches and personal interviews.
- 26.-27.05.2020, Virtual Retreat, Online (16 participants): Retreat of the doctoral researchers to network, get to know each other and discuss their projects and possible collaborations.
- 22.09.2020, General Assembly 2020, Online (30 participants): Assembly of the HIDSS4Health "parliament", consisting of all PIs, two postdocs and one selected doctoral researchers per cohort. Here, the current state of the school is discussed and members of the Steering Committee and new PI candidates are elected.
- 23.-24.09.2020, Retreat Bad Liebenzell (24 participants): Retreat of the doctoral researchers to network, get to know each other and discuss their projects and possible collaborations.
- 09.10.2020, Doctoral Researcher General Assembly 2020, Gastdozentenhaus des KIT (18 participants): Assembly of the doctoral researchers, where organizational aspects of the school are discussed and representatives for the General Assembly and the Steering Committee are elected.

In addition: Since July 2020 seven regular meetings of HIDSS4Health doctoral researchers (every 2-4 weeks).

Courses

- 13.-14.10.2020, Scientific Writing Course, KIT (10 participants, 8 HIDDS4Health, 1 Helmholtz, 1 external, 5 female): The seminar „Publishing in scientific journals“ covers all aspects of writing, submitting, reviewing and publishing a scientific manuscript.

3 Lectures as part of the Data Science & Health Lecture Series

(The first five lectures in 2020 were held at the Mathematikon of the University Heidelberg, the KIT and the DKFZ, later the lectures were given at Zoom. The number of participants varies from ten to approx. 50, and is significantly higher for the online lectures.)

- 07.01.2020, Introduction to Text Analysis (Michael Gertz)
- 21.01.2020, Requirements Engineering for Data-driven Solutions (Anne Koziolek)
- 21.01.2020, Introduction to Decentralized Data Management with Distributed Ledger Technology (Ali Sunyaev)
- 04.02.2020, Mathematical Aspects of Uncertainty Quantification (Martin Frank)
- 04.02.2020, Introduction to Visual Data Science (Filip Sadlo)
- 02.11.2020, Combinatorial Optimization Techniques for Bioimaging (Bogdan Savchynskyy)
- 02.11.2020, Medical Informatics in Translational Oncology (Frank Ückert)
- 16.11.2020, Time Series Analysis (Ralf Mikut)
- 16.11.2020, Data Inference on Sequence Data (Alexander Schug)
- 30.11.2020, Any Growth is Bounded – On the Future of Performance Scaling (Holger Fröning)
- 30.11.2020, The n<<p Paradigm in Omics Data Analysis (Benedikt Brors)
- 14.12.2020, Minimally-Invasive Robots for Medicine (Franziska Mathis-Ullrich)
- 14.12.2020, Mathematical Foundations of Deep Learning (Martin Frank)

2 Journal Club Series

Starting in February and October 2020, (4, resp. 5 regular participants, estimated): Discussion of different papers about machine learning. Takes place every second week.

Publications

A total of seven first author publications (journal/conference) were published and 13 co-authored publications (journal/conference).

First author publications (journal/conference):

1. Mikut, R. (2020), Machine Learning and Artificial Intelligence – a Revolution in Automation Technology or Only a Hype? at-Automatisierungstechnik, https://doi.org/10.1515/auto-2020-0041.
2. Ines Reinartz, Marie Weiel, Alexander Schug (2020), FRET Dyes Significantly Affect SAXS Intensities of Proteins, Israel Journal of Chemistry, https://doi.org/10.1002/ijch.202000007.
3. Stefan Haller, Mangal Prakash, Lisa Hutschenreiter, Tobias Pietzsch, Carsten Rother, Florian Jug, Paul Swoboda, Bogdan Savchynskyy (2020), A Primal-Dual Solver for Large-Scale Tracking-by-Assignmen, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics.
4. P. M. Scheikl, U. Scheler, N. Franke, and F. Mathis-Ullrich (2020), A Simulation Environment for Visual Multi-Agent Reinforcement Learning in Robot-Assisted Laparoscopy [Abstract], Biomedical engineering, vol. 65, no. s1.
5. P. M. Scheikl, J. Zhebin, and F. Mathis-Ullrich (2020), Path Planning for Robotic Camera Guidance in Laparoscopy [Abstract], Biomedical engineering, vol. 65, no. s1.
6. P. M. Scheikl, S. Laschewski, A. Kisilenko, T. Davitashvili, B. Müller, M. Capek, B. P. Müller-Stich, M. Wagner, and F. Mathis-Ullrich (2020), Deep Learning for Semantic Segmentation of Organs and Tissues in Laparoscopic Surgery, Current Directions in Biomedical Engineering, vol. 6, no. 1, https://doi/10.1515/cdbme-2020-0016.
7. C. Seibold, J. Kleesiek, HP. Schlemmer, R. Stiefelhagen (2020), Self-Guided Multiple Instance Learning for Weakly Supervised Thoracic Disease Classification and Localization in Chest Radiographs, ACCV2020.

Co-authored publications (journal/conference):

1. Scott Thiebes, Philipp A. Toussaint, Jaehyeon Ju, Jae-Hyeon Ahn, Kalle Lyytinen, Ali Sunyaev (2020), Valuable Genomes: Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing, Journal of Medical Internet Research, https://doi.org/10.2196/14890.
2. Scherr, T.; Löffler, K.; Böhland, M.; Mikut, R. (2020), Cell Segmentation and Tracking Using Cnn-Based Distance Predictions and a Graph-Based Matching Strategy, Plos One, https://doi.org/10.1371/journal.pone.0243219.
3. Jakob Rosenbauer, Chengting Zhang, Benjamin Mattes, Ines Reinartz, Kyle Wedgwood, Simone Schindler, Claude Sinner, Steffen Scholpp, Alexander Schug (2020), Modeling of Wnt-Mediated Tissue Patterning in Vertebrate Embryogenesis, PLOS Computational Biology, https://doi.org/10.1371/journal.pcbi.1007417.
4. Klaus Kades*, Jan Sellner*, Gregor Koehler, Peter M. Full, TY Emmy Lai, Jens Kleesiek*, Klaus H. Maier-Hein (2020), Adapting BERT to Assess Clinical Semantic Textual Similarity: Algorithm Development and Validation Study, JMIR Medical Informatics, 10.2196/22795
5. F. Mathis-Ullrich and P. M. Scheikl (2020), Robots in the Operating Room - (Co) Operation During Surgery, Gastroenterologe, https://doi.org/10.1007/s11377-020-00496-x
6. R. Bruch, P. M. Scheikl, R. Mikut, F. Loosli, and M. Reischl (2020), epiTracker: A Framework for Highly Reliable Particle Tracking for the Quantitative Analysis of Fish Movements in Tanks, SLAS TECHNOLOGY: Translating Life Sciences Innovation, https://doi/10.1177/2472630320977454.
7. Fabian Isensee, Paul F. Jäger, Simon A. A. Kohl, Jens Petersen, Klaus H. Maier-Hein (2020), nnU-NetA Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation, Nature Methods, https://doi/10.1038/s41592-020-01008-z
8. Gianluca Brugnara, Fabian Isensee, Ulf Neuberger, David Bonekamp, Jens Petersen, Ricarda Diem, Brigitte Wildemann, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus H. Maier-Hein, Philipp Kickingereder (2020), Automated Volumetric Assessment With Artificial Neural Networks Might Enable a More Accu-Rate Assessment of Disease Burden in Patients With Multiple Sclerosis, European Radiology, https://doi/10.1007/s00330-019-06593-y.
9. Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M. Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Anna Kisilenko, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Minu Tizabi, Matthias Eisenmann, Tim J. Adler, Janek Gröhl, Melanie Schellenberg, Silvia Seidlitz, TY. Lai, Veith Roethlingshoefer, Fabian Both, Sebastian Bittel, Marc Mengler, Martin Apitz, Stefanie Speidel, Hannes G. Kenngott, Beat P. Müller-Stich (2020), Heidelberg Colorectal Data Set for Surgical Data Science in the Sensor Operating Room, arXiv preprint https://arxiv.org/abs/2005.03501.
10. Leonardo Ayala, Silvia Seidlitz, Anant Vemuri, Sebastian J. Wirkert, Thomas Kirchner, Tim J. Adler, Christina Engels, Dogu Teber, Lena Maier-Hein(2020), Light Source Calibration for Multispectral Imaging in Surgery, International Journal of Computer Assisted Radiology and Surgery.

11. Janek Gröhl, Melanie Schellenberg, Kris Dreher, Lena Maier-Hein (2020), Deep Learning for Biomedical Photoacoustic Imaging: A Review, arXiv, preprint https://arXiv:2011.02744.
12. Niklas Holzwarth, Melanie Schellenberg, Janek Gröhl, Kris Dreher, Jan-Hinrich Nölke, Alexander Seitel, Minu D. Tizabi, Beat P. Müller-Stich, Lena Maier-Hein (2020), Tattoo Tomography: Freehand 3d Photoacoustic Image Reconstruction With an Optical Pattern, arXiv preprint https://arXiv:2011.04997.
13. G. Salg, M. Ganten, A. Bucher, H. Kenngott, M. Fink, C. Seibold, R. Fischbach, K. Schlamp, C. Velandia, P. Fervers, F. Doellinger, A. Luger, S. Afat, U Merle, M Diener, P Pereira, T Penzkofer, T. Persigehl, A. Othman, C. Heussel, M. Baumhauer, G. Widmann, K Stathopoulos, B Hamm, T. Vogl, K. Nikolaou, H. Kauczor, J. Kleesiek (2020), An International Feasibility Study Using a Reporting and Analysis Framework for the Structured Evaluation of COVID-19 Clinical and Imaging Data, Nature Digital Medicine.

Further news from HIDSS4Health

- The School was present at the following conferences or events:
- 2020, 54th-annual-conference-of-the-german-society-for-biomedical-engineering (BMT), oral presentation
- March 2020, Bildverarbeitung für die Medizin, oral presentation
- 03.-07.04.2020, Workshop: ISBI Cell Tracking Challenge at the IEEE International Symposium on Biomedical Imaging (ISBI 2020), Iowa City, IA, USA (virtual), oral presentation
- 17.-20.08.2020, Society for Mathematical Biology (SMB) 2020 Annual Meeting, poster presentation
- 27.08.2020, Artificial Intelligence and Statistics 2020, oral presentation
- 17.-19.09.2020, 19. Jahrestagung der Deutschen Gesellschaft für Computer- und Roboterassistierte Chirurgie e.V. (CURAC), oral presentation
- 19.10.2020, Medical Image Computing and Computer Assisted Intervention (MICCAI), poster presentation
- 30.11.2020, ACCV2020 (Asian Conference on Computer Vision), poster presentation
- 12.12.2020, NeurIPS (Medical Imaging meets NeurIPS workshop), poster presentation

The following prizes were awarded to the school's members:

- Best reviewer award Workshop on Security and Privacy 2020; PI: Ali Sunyaev, DR: Mikael Beyene
- IEEE ISBI Cell Tracking Challenge: 2x First Rank, 2x Second Rank, 5x Third Rank, PI: Ralf Mikut; DR: Katharina Löffler
- Cooperation with Fraunhofer IPA (Mannheim) in the context of reinforcement learning for autonomous catheters (new partners), PI: Mathis-Ullrich, DR: P.M. Scheikl
- 1st place at HIDA DATATHON FOR GRAND CHALLENGES ON CLIMATE CHANGE, PI: L. Maier-Hein, DR: Melanie Schellenberg & Patrick Scholz
- MICCAI 2020 Best Challenge Reviewer Award, PI: L. Maier-Hein, DR: Patrick Scholz

Communication/Marketing

HIDSS4Health has a **website** (www.hidss4health.de), a **Twitter account** with 304 followers (feb 2021), and a **mailing list**.

PAULA BREITLING – HIDSS4HEALTH

Software Assistant for Oncologists

Oncologists who want to provide cancer patients with tailored treatment have to analyze large amounts of information – which takes a lot of time and expertise. At the German Cancer Research Center, Paula Breitling contributes to the development of a software that aims to help doctors recommend appropriate therapies. This could allow more people to benefit from custom-fit treatment

"No dissecting, no blood!" – this was one of Paula Breitling's conditions when she chose her field of study. For her, data and information were the abstract objects that she found more fascinating – despite having an avid interest in life science topics. Now, the information specialist is demonstrating in her doctoral thesis that it's also possible to save lives with data expertise rather than swabs or scalpels. Breitling wants her research to support doctors with the difficult task of providing suitable therapy recommendations for the treatment of cancer patients and offer more people access to personalized tumor therapy in the process.

She is a doctoral researcher at the Helmholtz Information & Data Science School for Health (HIDSS4Health), which is part of the Helmholtz Information and Data Science Academy (HIDA) – Germany's largest postgraduate training network in the information and data sciences. HIDSS4Health offers Breitling a unique mix of data science and medical research that is vital to a project like hers. Abstract data come together with a very tangible field of application here – oncology. "Cancer is an issue that affects many people – even if it is their relatives and friends who get sick," says Breitling. She herself had direct experience with this while completing her degree, as one of her fellow students was diagnosed with cancer while they were preparing for their exams together. "If I am able to use my data science background to do something that helps people who are affected by cancer, I would find that very satisfying."

Breitling's expertise in applying data science methods to structure large quantities of data comes into play where standard therapy approaches in cancer treatment are no longer effective and the doctors treating the patients have to seek other potential therapies. To do this, they need to assess complex information relating to the respective patients, because the biological characteristics of a tumor disease don't just vary between different types of cancer but also from patient to patient.

A lot of information must be analyzed for a therapy recommendation

The first step in identifying other potential therapy approaches is sequencing the patients' DNA. Looking at the genome makes it possible to see whether there are specific mutations associated with cancer. This can offer an indication as to whether certain combinations of drugs can be used. Oncologists provide therapy recommendations based on this data and other information relevant to the patient. Breitling says that experience plays an important part in this: "Some oncologists already know what direction they need to move in if they find certain characteristics; doctors who are relatively new to this field, on the other hand, have to spend significantly more time searching and verifying and have to throw out a lot of information." Even an experienced doctor would need around two hours to complete this work, Breitling says, while it would take well over a day for a young colleague to put together a sound recommendation for just one patient. Making this process simpler for all doctors is the goal that Breitling is pursuing alongside other researchers. Her job is to capture the workflow that a doctor negotiates while making their decision, so she can then model this workflow with the help of a computer using data science methods and map it in a software solution.

This software – the Knowledge Connector – is designed to connect all these different units of knowledge and is being developed at the German Cancer Research Center (DKFZ) in Heidelberg. Breitling's own research project is split between this location, where she is a member of Frank Ückert's team, and the Karlsruhe Institute of Technology (KIT), where she works on the other half of her project, data science, in Michael Beigl's research group.

"Clicking together" the relevant information by using a software

The Knowledge Connector links various external databases and gives the doctor a clear overview of the collected information with respect to the individual characteristics of a patient and their tumor. Information on genetic variants, genome mutations, drugs, and various clinical studies and publications are incorporated into this evaluation process – a multifaceted task that doctors have had to carry out on their own to date. Based on all this information, it then becomes easier for the specialist to put together a ranking of promising treatment variations. And the tool can also highlight suitable drug studies that are currently underway and could accommodate the patients, for example.

The software doesn't put together any therapy recommendations itself – but, "ideally, applying this software would give the oncologist all the information they need to make their decision, and they wouldn't have to do all the research themselves anymore," Breitling explains. The advantage of this type of tool is that it merges clinical and scientific data that is usually documented in different IT systems and isn't generally suitable for being exchanged or analyzed in conjunction with other data. Using the new tool, doctors should be able to put together a list of the therapy recommendations based on the parameters that are deemed important and transfer them directly into the appropriate text form. This eliminates a lot of the work involved in copying and writing information, which eats up doctors' time. The benefit is obvious: the doctors have a better overview, there aren't so many different file formats, and a therapy recommendation can be put together more quickly. By making the workflow simpler, the software can ultimately play a role in helping more people benefit from personalized (and therefore more effective) treatment than previously.

"Understanding a doctor's entire knowledge process"

A prototype of the Knowledge Connector is already in the initial testing phase. But the system is still far from being able to do everything doctors would need. Breitling is addressing this by looking at exactly what they do with the program: "We need to understand a doctor's entire knowledge process," she says. "In order to make an algorithm on this basis, I need to know, for example, how often they click, how often they switch to a different page, when text entered was where, and what search steps were carried out at what point in time." It goes without saying that this isn't done by peering over the doctor's shoulder; these processes are recorded by machine. A simple piece of software that runs alongside the program stores data regarding keystrokes, clicks, and the websites accessed; it records the time spent on a website, text that is keyed in, and many other parameters. Vast quantities of data are created in the process. Breitling has yet to make any final decisions on the methods she will use to subsequently model these processes – but the young researcher can already pick up on certain weaknesses in the software when she takes a cursory glance at the vast datasets in her first test sequence. Her ongoing analysis will focus on the following questions: How do doctors conduct their research right now? How do they use the software? What other features does the Knowledge Connector need, and how should the user interface be designed? A total of nine oncologists will test the software at a later stage in the project – at least, this is what Breitling fervently hopes for, so she can soon have a broader database to work with.

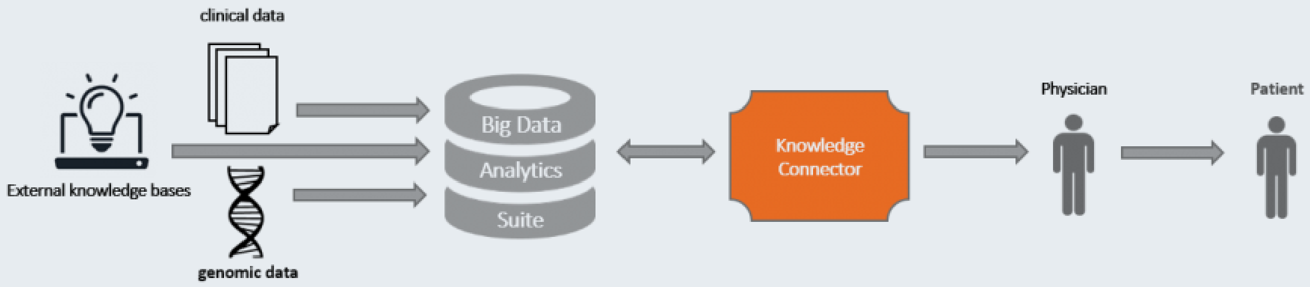
Designing such an extensive study and organizing a multifaceted research process sounds challenging, but the eloquent doctoral researcher doesn't find this daunting. For Breitling, even the demanding selection procedure that HIDSS4Health set out for its applicants motivated her rather than scaring her off: "I just thought, I'll give it a try!" Out of nearly 220 applicants, eleven candidates were ultimately selected for the first cohort at HIDDS4Health – and Breitling was one of them. She is now happy to have the opportunity to benefit from an inspiring network and is taking her involvement a step further: As the spokesperson for the doctoral researchers, she represents their interests in the steering committee – and therefore also plays a small part in shaping the ongoing development of HIDSS4Health.

As far as her own research is concerned, Breitling is looking eagerly ahead to the moment when she finally has all the data she needs. Her enthusiasm is evident when she talks about her work: "Once you've actually got the data, you can do wonderful things with it!" For example, processing it in a way that ultimately makes it possible for more people than ever before to have access to cancer treatment tailored specifically to them. It's medical assistance thanks to data science – without scalpels, but with a lot of heart and soul.

Author: Constanze Fröhlich

https://www.helmholtz-hida.de/en/activities/news/newsdetail/software-assistent-fuer-onkologen/

Credits: Paula Breitling: private/Schematic function of the Knowledge Connector: DKFZ



ALEXANDRA WALTER – HIDSS4HEALTH

Fighting Cancer with Artificial Intelligence

How can diseased tumor tissue be destroyed in a targeted manner while preserving the healthy tissue immediately surrounding it? Alexandra Walter uses data science and artificial intelligence to make radiation therapy even more precise

When people are diagnosed with cancer, they have to face a difficult fight that literally takes everything they've got – because radiation not only destroys tumor tissue, it always destroys a certain amount of healthy tissue as well. The only question is how much healthy tissue is subjected to radiation and how extensive the sacrifice is. Healthy tissue damaged by radiation often loses some of its function afterwards too. One of the most common side effects of radiation in the mouth and neck region is dryness of the mouth caused by the salivary glands being impaired. This can be very unpleasant for those affected.

Alexandra Walter is committed to keeping all these negative impacts of radiation to a minimum. For her doctoral thesis, the 26-year-old is working to pinpoint malignant tumors in the head and neck regions of cancer patients with enough precision for diseased tissue to be destroyed as completely as possible while at the same time leaving as much healthy tissue as possible untouched.

More diseased tissue could be destroyed, and more healthy tissue could be preserved

Alexandra Walter is committed to keeping all these negative impacts of radiation to a minimum. For her doctoral thesis, the 26-year-old is working to pinpoint malignant tumors in the head and neck regions of cancer patients with enough precision for diseased tissue to be destroyed as completely as possible while at the same time leaving as much healthy tissue as possible untouched.

Currently, doctors need to mark the tissue to be targeted by radiation on CT images prior to treatment. "This can take up to three hours if you don't have any other tools. The algorithms already available to doctors use another patient's data as the starting point," says Walter. She notes that this approach does work to a certain extent, but only to a certain extent. Walter is convinced that it can be significantly improved upon. More diseased tissue could be destroyed, and more healthy tissue could be preserved. This would benefit many, many people, given that head and neck tumors are the fourth most common type of cancer, with more than 20,000 new cases in Germany every year.

In the first phase of her PhD at the Helmholtz Information and Data Science School for Health (HIDSS4Health), Walter is initially working to improve an existing, intelligent computer program that independently detects tumor tissue. The program then provides the doctor with an immediate suggestion as to where to direct the radiation. The program is referred to as an artificial neural network, a highly complex software system that uses extensive internal circuitry with a structure similar to that of the neural network in the human brain to learn independently, so to speak. Because of this, the system is also described as a type of artificial intelligence. "The program is, in a sense, trained, because it incorporates a wide range of data from completed courses of radiation therapy. I want to improve the program by adapting its structure and adding new ways of learning. The ultimate goal is to make it much more precise," Walter explains. This would mean improved prognoses for patients, and subsequent impacts such as damage to the larynx and the associated speech impairment would become less common.

Walter's approach consists of testing, combining, calculating, and programming; a mix of mathematics and computer science. This is the exact combination she learned while studying for her bachelor's degree in cognitive science at the University of Tübingen. In addition to neurobiology, this field incorporates computer science and mathematics in particular. Walter followed this with a master's degree that focused on the field of logic and wrote her thesis on the interpretability of neural networks. She is now able to conduct research at HIDSS4Health, which is part of the Helmholtz Information and Data Science Academy (HIDA), Germany's largest postgraduate training network in information and data science. Walter enjoys this not only because she can apply many of the things she has been focusing on over the past few years. "I love how interdisciplinary it is here. Even though it has become something of a buzzword, the interplay between multiple disciplines is in fact an everyday reality in my PhD," Walter says.

What is cancer, how does it occur and why is it so difficult to treat?

Even before she moved to Karlsruhe about six months ago to start her doctorate at HIDSS4Health, she bought and worked her way through a number of medical textbooks. "Medicine is the component I addressed least during my degree. So that was where I needed to invest most energy in the beginning," Walter says. She is now familiar with the anatomy of the head and neck and understands what cancer is, how it occurs, and why it's so difficult to treat.

This knowledge was then quickly reinforced at HIDSS4Health by a combination of practical experience, opportunities to exchange ideas, and explanation. One key factor was simply the research institutions involved in the program – the Karlsruhe Institute of Technology (KIT), Heidelberg University, and the German Cancer Research Center (DKFZ). Like these institutions, HIDSS4Health is also based in Heidelberg and Karlsruhe.

All doctoral researchers on the program, including Walter, have two supervisors from two disciplines. In her case, one is from the field of health science at DKFZ, and one is from data science at KIT. And then there are the lectures, workshops, seminars, and summer schools. “Most of these events bridge the gap between health science and data science. Given my background, my starting point is the data science side of things, and my colleagues with medical backgrounds contribute to my work with their knowledge of health science,” explains Walter.

The researchers come together, discuss their work and are creating new knowledge in the process

Knowledge is the key currency in an environment like this, where experts from different areas work together and learn from each other. Everyone knows something different and the researchers come together, support each other, exchange ideas and experiences, and discuss their work, creating new knowledge in the process. Walter thinks it’s particularly useful to have everyone look at the same problem from different perspectives. In many cases, each new point of view yields new insights, too. “If I look at neural networks through the lens of computer science, for example, they are essentially tools. If I look at them from the perspective of mathematics, they are mathematical functions. And if I combine the two, I can better understand these tools and improve them. It might sound strange, but this sort of thing makes me really happy,” Walter smiles.

Walter’s résumé would suggest that she’s been working towards her doctoral post at HIDSS4Health for years. But that wasn’t the case. “I didn’t have a specific goal. I just always went in whatever direction I found interesting,” Walter says. It was only when she wrote her application for the doctoral post at the Graduate School that she realized she would fit in really well there! The path that Walter has taken is a perfect example of how a certain amount of inexperience and open-mindedness can sometimes be the right approach. Do what interests you and success will follow on its own. “It’s worked for me so far, and I’m thankful for that,” she says.

Walter’s thesis can truly benefit patients, and this also motivates her. “I’ve made a few visits to the centers where patients go for radiation. The fact that my work can help to prolong their lives and improve their chances of recovery makes everything I do even more important, and I also feel a certain sense of responsibility.”

As a neural network, the program learns with each correction

Walter will also need motivation and a sense of responsibility for what lies ahead over the next two and a half years. After the first phase of her PhD, which involves improving the current assistance program used for radiation therapy, Walter will try to integrate doctors’ knowledge and experience into the neural network. Among other things, this will include practical knowledge and existing guidelines for radiotherapists. “The program will never be able to replace doctors. But ideally, it will be able to determine where the radiation should be focused, so the doctor will only have to approve it or, at most, make small adjustments,” says Walter.

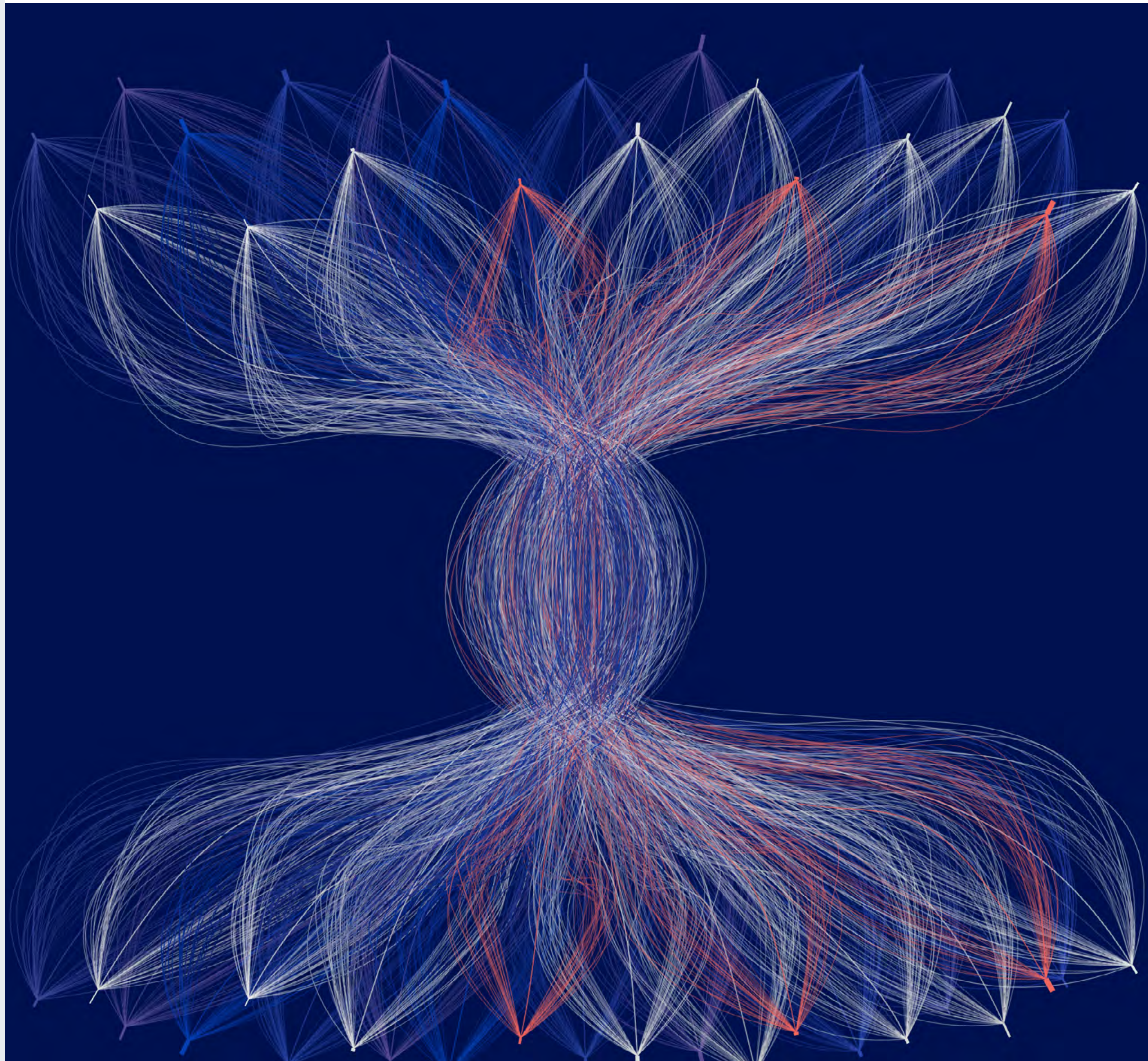
Once the program is completed, the third phase will focus on optimizing it. “As a neural network, the program essentially learns from every correction a doctor makes to it. But, of course, I want to get everything done before it’s applied in clinical practice in order to ensure a high level of precision from the start,” Walter explains. During the process, the focus is always on two, closely interlinked aspects: Where should the radiation be directed to specifically destroy bad, diseased tissue, and what should it avoid where possible to preserve good, healthy tissue in a targeted manner. This is Walter’s mission.

Text: Christian Heinrich

https://www.helmholtz-hida.de/en/activities/news/newsdetail/fighting-cancer-with-artificial-intelligence/

Credits: Alexandra Walter: private/Head matter: Unsplash

HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY



CHRISTIAN GERLOFF (HDS-LEE) applies artificial intelligence and machine learning methods to the neurosciences to gain a better understanding of the human brain.

The dataset on which this picture is based contains signals from 22 different regions of the human brain. The respective data come from two participants who were interacting with each other as part of an experiment and relate to the brain region, the source, the measuring point coordinators, and the rate at which concentrations of oxyhemoglobin change.

HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE, EARTH AND ENERGY (HDS-LEE)

The HDS-LEE structured doctoral program aims at excellent graduates of mathematics, computer science, natural sciences and engineering from all over the world. The doctoral researchers at HDS-LEE are trained in all essential areas of Information and Data Sciences as well as in communication and other key qualifications. The training components of the program are strengthened by individually tailored training measures, e.g. at the Jülich Supercomputing Center (JSC).

Research areas:
Life Sciences, Earth Science and Energy Systems resp. Material Sciences.

Partners of HDS-LEE:
RWTH Aachen University, University of Cologne, Max Planck Institute for Iron Research, German Aerospace Center and Jülich Research Center

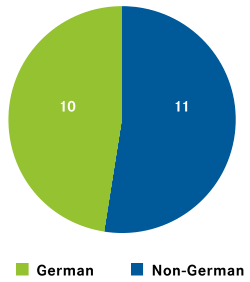


Fig. 20: Recruitment of doctoral researchers from abroad at HDS-LEE 2019-2020 (excl. associated doctoral researchers).

Applicant situation and recruitment

While 13 positions were filled in 2019, eight new PhD positions were awarded in 2020. In 2019, a total of 330 people from 70 nations applied; in the following year, 90 applications came in from 28 nations. In both years, the high proportion of applicants from Iran, India and China should be emphasized.

The current 21 PhD positions at HDS-LEE are held by people from 11 nations, ten of them are Germans. Three women are among the doctoral researchers. Ten other doctoral researchers are associate members, including three women. Of the associated researchers, eight belong to FZ Jülich and two to RWTH Aachen University.

Events and networking

While there were only two HDS-LEE events in 2019 (one networking event and one lecture), in 2020 significantly more events took place, namely 41 (9 networking events, 16 lectures, 16 courses).

› Networking events

4 Doctoral Seminars (04.06. & 09.06., 22.06. & 26.06.2020): The doctoral seminar was created as a compensation for the retreat, since that could not take place on-site in 2020. The annual internal retreat is designed as a platform where doctoral students inform each other about their latest research results and open issues. It is a forum for knowledge transfer and information exchange. It follows a Docs-for-Docs concept, where you select content and conference style. The doctoral students give presentations on current research topics (both results and open issues). The coordinator team organizes the retreat, which contains social program incl. institute visits. https://www.hds-lee.de/events/internal-retreat/, HDS-LEE intern, 37-41 participants

3 Monthly Doctoral seminars: Discussion groups on video presentations (16.09., 21.10., 25.11.2020): The doctoral seminar was created as a compensation for the retreat, since that could not take place on-site in 2020. In the discussion group meeting, one doctoral student presents their work and answers questions. The aim is to exchange knowledge and build cooperation between the doctoral students, HDS-LEE intern, 15-24 participants

Weekly Coffee break (since 15.04.2020): The monthly virtual coffee breaks were initiated to enable an informal exchange between the doctoral students and to have the possibility to have regular contact to the coordinator team, HDS-LEE intern, 5-10 participants

Virtual Christmas Event and Election of PhD spokesperson (16.12.2020): In the Christmas event, online games were played for social networking, HDS-LEE intern, 20-25 participants

› Courses

1 Node-Level Performance Engineering Course (20-22.01.2020), Cologne (CDS at UoC): This course covers performance engineering approaches on the compute node level. Even application developers who are fluent in OpenMP and MPI often lack a good grasp of how much performance could at best be achieved by their code. https://cds.uni-koeln.de/index.php?id=14559, participants from UoC, RWTH, FZJ

6 EU Regional School Courses

The EU Regional School is open to graduate students from universities in Germany, Belgium, and the Netherlands. There are 5 courses per term. It is a series of three-hour "short courses“ on introductory and advanced various topics in computational science, and are attended by a cross-section of graduate students in computational science (fluid mechanics, computer science, molecular dynamics, contact mechanics, reduced-order modeling, etc.). Organized by AICES, RWTH and JARA-CSD (missing course sessions were postponed due to Corona situation), https://blog.rwth-aachen.de/irtg-mip/eu-regional-school-videos-2020-part-1/

- 24.01.2020 (Course 1, RWTH), Prof. Dr. Siddhartha Mishra: Current Topics in Numerical Methods for Hyperbolic Systems of Conservation Laws: Uncertainty Quantification, Statistical Solutions and Machine Learning, participants 30-40
- 15.07.2020 (Course 5), Prof. Dr. Thomas Böhlke: Integrated Engineering of Continuous-Discontinuous Long Fiber Reinforced Polymer Structures, 30-40 participants, https://blog.rwth-aachen.de/irtg-mip/event/eu-regional-school-with-prof-dr-thomas-boehlke/
- 17.07.2020 (Course 7), Prof. Dr. Raul F. Tempone : Forward and Inverse Problems with Multilevel Monte Carlo, 30-40 participants, https://blog.rwth-aachen.de/irtg-mip/event/eu-regional-school-with-prof-dr-raul-f-tempone/
- 28.07.2020 (Course 89, Prof. Dr. Felix Krahmer: Structure and Randomness in Data Science, 30-40 participants, https://blog.rwth-aachen.de/irtg-mip/event/eu-regional-school-with-prof-dr-felix-krahmer/
- 04.08.2020 (Course 6), Prof. Leszek Demkowicz, Ph.D.: The Discontinuous Petrov-Galerkin (DPG) Method (with Optimal Test Functions), 30-40 participants, https://blog.rwth-aachen.de/irtg-mip/eu-regional-school-videos-2020-part-2/
- 15.12.2020 (Course 9), Dr. Leonardo Uida, Ph.D.: Geophysical Forward and Inverse Modeling, 30-40 participants, https://blog.rwth-aachen.de/irtg-mip/event/eu-regional-school-with-dr-leonardo-uida-ph-d/

4 Lectures on Data Science, Methods & Applications

The Lectures on Data Science, Methods & Applications aim to teach and train the doctoral students in all essential elements of data science and information. The doctoral students are introduced to the respective disciplines and data science methods, get an insight and apply what they have learned in hands-on sessions. https://www.hds-lee.de/events/lectures-on-data-science/

- 25.06.2020, Data Sources in Biology Systems, Neuroscience, Energy Systems, and Geoscience, 27 participants from HDS-LEE
- 02.07.2020, Data Science, Data Management and Scientific Workflows, 27 participants from HDS-LEE
- 04.11.2020, Machine Learning in a Nutshell Part 1: Neuronal Networks/ DL, 29 participants from HDS-LEE, 5 from RWTH & FZJ
- 13.11.2020, Machine Learning in a Nutshell Part 2: Hyperparameter Optimization, 31 participants from HDS-LEE, 24 external

Transferable Skill Courses

Transferable skills courses aim to foster the professional skills and personal development essential to the successful completion of doctoral projects. Transferable Skill Courses are offered to the entire group of HDS-LEE doctoral students and are organized via JuDocs, a center for transferable skills training for doctoral researchers at FZJ.

- 27.03.2020: Good scientific practice, 20 participants from HDS-LEE
- 18.-19.06. & 16.-17.07.2020: Doing Science 1&2, 11-12 participants from HDS-LEE, https://www.fz-juelich.de/judocs/EN/structured_doctoral_support/TransferableSkillsCourses/1_Mandatory_Courses/2_Doing_Science/_node.html
- 27.-28.08. & 25.-26.11.2020: Scientific writing 1&2, 12 participants from HDS-LEE, https://www.fz-juelich.de/judocs/EN/structured_doctoral_support/TransferableSkillsCourses/1_Mandatory_Courses/1_Scientific_Writing/_node.html

- **Lectures**

5 HDS-LEE Seminar Series

In the “HDS-LEE Seminar Series”, data science experts from the application areas Life, Earth and Energy are invited to present the current state of their research to interested researchers. The goal is that the doctoral students get to know the experts in the individual fields and have the opportunity to interact and discuss with them (each lecture with 30-40 participants from HDS-LEE, RWTH, FZJ, UoC, MPIE, DLR). https://www.hds-lee.de/events/seminar-series

- 18.12.2019, Prof. Ph.D. Yannis Kevrekidis: No Equations, no Variables, no Parameters: Data and the Computational Modeling of Complex/Multiscale Systems
- 08.01.2020, Jun-Prof. Dr. Julia Kowalski: Digital Earth Science – Predictive Simulation in Natural Hazards Research
- 15.01.2020, Dr. Eva-Maria Gerstner: Research Data Management – A Short Introduction into the Subject and the Central Services of ZB
- 08.06.2020, Dr. Maria Vittoria Barbarossa - Mathematical Models in Epidemiology
- 02.12.2020, Laura Helleckes and Michael Osthege: Productionizing Bayesian Nowcasting Workflows with Apache Airflow

2 Charlemagne Distinguished Lecture Series 2020

Twice a year the AICES doctoral students organize the prestigious Charlemagne Distinguished Lecture Series as part of the SSD Seminar Series, whose objective is to invite persons, who have achieved impressive accomplishments throughout their career and, in this sense, to get inspired by their scientific achievements (each lecture with 30-40 participants from RWTH, HDS-LEE, FZJ)

- 31.01.2020, Prof. Anthony T. Patera, Ph.D.: Parametrized Partial Differential Equations: Mathematical Models, Computational Methods, and Applications, Charlemagne Distinguished Lecture Series Videos 2020 Part 1 «IRTG Modern Inverse Problems (MIP)« (rwth-aachen.de)
- 19.10.2020, Prof. George Em Karniadakis, Ph.D. – Physics-Informed Neural Networks (PINNs): An Alphabet of Algorithms for Diverse Applications, Charlemagne Distinguished Lecture Series Video 2020 Part 2 «IRTG Modern Inverse Problems (MIP)« (rwth-aachen.de)

8 SSD Seminar Series

Experts are invited to present their research in a seminar series, operating on a semester-based schedule. The aim is to invite speakers relevant to the SSD, but also potentially of interest to a wider audience. Such seminars provide opportunities for learning about state-of-the-art research, and for interaction and discussion with top experts. The seminar series invites up to 20 speakers per year. A combination of national and international guest speakers is anticipated. https://blog.rwth-aachen.de/irtg-mip/ssd-seminar-series-2020-part-1/

- 15.06.2020, Prof. Sebastian Krumscheid, Ph.D.: Beyond Multilevel Monte Carlo Methods for Expected Values
- 22.06.2020, Prof. Dr. Herbert Egger: On Model Order Reduction for Inverse Problems in Tomographic Imaging Applications
- 29.06.2020, Dr. Marc S. Boxberg: A Geophysicist’s Perspective on the Modeling of Mechanical Waves
- 06.07.2020, Prof. Dr. Holger Gohlke: From Protein Structure Prediction to Mechanistic Insights into Signaling and Disease Diagnostics
- 13.07.2020, Giulia Rossetti: Computational Approaches To Drug Repurposing and Design in COVID-19 Pandemic
- 02.11.2020, Prof. Andrea Saltelli, Ph.D.: Composite Indicators: Lights and Shadows
- 09.11.2020, SSD Seminar Series with Prof. Ramon Codina, Ph.D.: Reduced Order Models for Flow Problems: Stabilization and Accuracy Enhancement
- 07.12.2020, Dr. Dominik Bongartz: Accelerating Global Optimization for Engineering Design via Reduced-Space Formulations and Tailored Relaxations

In the **Seminar Series**, data science experts from the application areas Life, Earth and Energy are invited to present the current state of their research to interested researchers. The goal is that the doctoral students get to know the experts in the individual fields and have the opportunity to interact and discuss with them (30-40 participants from RWTH, HDS-LEE, FZJ). https://indico-jsc.fz-juelich.de/event/140/

- 06.07.2020, Blake Richards: Predictive Cost Functions in the Neocortex

Publications

A total of eight first author publications (journal/conference) and two co-authored publications (journal/conference) were published in 2020, six first author publications are in preparation.

Published first author publications:

1. L. R. Gorjão, M. Anvari, H. Kantz, C. Beck, D. Witthaut, M. Timme, B. Schäfer (2020), Data-driven Model of the Power-grid Frequency Dynamics, IEEE Access 8, 43082 [DOI] [ArXiv].
2. L. R. Gorjão, R. Jumar, H. Maass, V. Hagenmeyer, G. C. Yalcin, J. Kruse, M. Timme, C. Beck, D. Witthaut, B. Schäfer (2020), Open Database Analysis of Scaling and Spatio-temporal Properties of Power Grid Frequencies, Nature Communications 11, 6362.
3. A. Yegenoglu et al. Ensemble Kalman (2020), Filter Optimizing Deep Neural Networks: An Alternative Approach to Non-Performing Gradient Descent, The 6th International Conference on Machine Learning, Optimization, and Data Science (LOD), LNCS Springer.
4. D. T. Doncevic, A.M Schweidtmann, Y. Vaupel, P. Schäfer, A. Caspari, A. Mitsos (2020), Deterministic Global Nonlinear Model Predictive Control With Recurrent Neural Networks Embedded, In: [IFAC World Congress 2020, Ifac 2020, 2020-07-11 - 2020-07-17, Berlin, Germany], in press.
5. J. F. Jadebeck, A. Theorell, S. Leweke, K. Nöh (2020), HOPS: High-Performance Library for (Non-) Uniform Sampling of Convex-Constrained Models, *Bioinformatics*, btaa872.
6. J. Kruse, B. Schäfer, D. Witthaut (2020), Predictability of Power Grid Frequency, IEEE Access 8, 149435-149446.
7. J. Kruse, B. Schäfer, D. Witthaut (2020), Pre-Processed Power Grid Frequency Time Series, Zenodo, https://doi.org/10.5281/zenodo.3744121.
8. M. Rüttgers, S.-R. Koh, J. Jitsev, W. Schröder (2020), A. Lintermann, Prediction of Acoustic Fields Using a Lattice-Boltzmann Method and Deep Learning BT - High Performance Computing, Proceedings of the 35th International Conference, ISC High Performance.

First author publications, in preparation

1. L. Boledi, S. Elgeti and S. Perotto, Casting Design Optimization via Hierarchical Model Education. Elsevier Applied Numerical Mathematics, in preparation.
2. A. Simson, H. Löwe, and J. Kowalski, Elements of Future Snowpack Modeling - Part 2: A Modular and Extendable Eulerian-Lagrangian Numerical Scheme for Coupling Transport, Phase Changes and Mechanics, in preparation.
3. Santarpia et al., Olfaction DB: A Database of Olfactory Receptor-Odorant Pairs, in preparation
4. L. Boledi, B. Terschanski, S. Elgeti and J. Kowalski, A Space-Time Fe Level-Set Method for Phase-Change Processes, Journal of Computational Physics, in preparation.
5. M. Samadi, S. Kiefer, A. Schuppert, A Learning Strategy for Hybrid Mechanistic/Data-Driven Models to Break the Curse of Dimensionality, in preparation.
6. M. Samadi, K. Sharafutdinov, A. Schuppert, Mechanisms of Disease Progression Within a Cohort of Severely Ill Patients Consisted of COVID-19 Cases, in preparation.

Co-authored publications

1. S. Fritsch, K. Sharafutdinov, M. Samadi, G. Marx, A. Schuppert, J. Bickenbach, Development and Validation of a Model Predicting Mortality Risk in Mechanically Ventilated COVID-19 Patients Requiring Intensive Care Treatment, Science Advances, submitted
2. A. Theorell, J. F. Jadebeck, K. Nöh, J. Stelling, PolyRound: Polytope Rounding for Random Sampling in Metabolic Network, in Preparation

Further news from HDS-LEE

The School was present at the following conferences or events:

- Doncevic, Schweidtmann, Vaupel, Schäfer, Caspari, Mitsos, Deterministic Global Nonlinear Model Predictive Control With Recurrent Neural Networks Embedded: 21st IFA World Congress, Germany, July 11-17, 2020, oral presentation.
- J. Kruse, Predictability of Power Grid Frequency, Conference on Complex Systems 2020.
- M. Rüttgers, S.-R. Koh, J. Jitsev, W. Schröder, A. Lintermann, Prediction of Acoustic Fields Using a Lattice-Boltzmann Method and Deep Learning BT - High Performance Computing, Proceedings of the 35th International Conference, ISC High Performance 2020, oral presentation.
- S. Malzacher, Standardised Data Acquisition in Biocatalysis According to the Fair Data Principles, Workshop Systematic profiling of multicopper oxidases 10/2020, oral presentation.
- S. Malzacher, J. Range, C. Halupczak, J. Pleiss, D. Rother, BioCatHub, BioCatHub, An Online Platform for Standardised Data Acquisition in Biocatalysis, Conference Amine biocat University of Stuttgart 02/2020, poster presentation.
- S. Malzacher, J. Range, C. Halupczak, J. Pleiss, D. Rother, BioCatHub, A Graphical User Interface for Standardized Data Acquisition in Biocatalysis, 10. ProcessNet Jahrestagung, Aachen, Germany, 21 Sep 2020 - 24 Sep 2020, poster presentation.
- A. Simson, Numerical Solution of the Mass Continuity Equation for Snowpack Modeling on Moving Mesh-ES: Coupling Between Mechanical Settling and Water Vapor Transport, EGU 2020, Online Conference, poster presentation.
- Furthermore, HDS-LEE was represented with oral presentations at the International Conference on Machine Learning and AI in (bio)Chemical Engineering.
- Anna Simson, Lisa Beumer, Hu Zhao and Leonardo Rydin Gorjao have successfully applied to the HIDA Trainee Network.
- Christian Gerloff will participate in the HIDA Summer Exchange Program with DSRC@BGU in Be’er Sheva Israel.
- Laura Helleckes received the Prize of the DECHEMA Biotechnology Future Forum in 2020.
- Prof. Alexander Mitsos successfully applied for funding by the DFG priority program with the title “Machine Learning in Chemical Engineering - Knowledge Meets Data: Interpretability, Extrapolation, Reliability, Trust”.
- HDS-LEE PI Priv.-Doz. Dr. Julia Kowalski has been included in the DFG Heisenberg Program.
- HDS-LEE PI Priv.-Doz. Dr. Julia Kowalski successfully acquired the BMU “KI Leuchtturmprojekt” (AI Lighthouse Project) with the title “KI:STE- KI-Strategie für Erdsystemdaten” (AI strategy for Earth system data).

Communication/Marketing

HDS-LEE has a **website** (www.hds-lee.de) and a **Twitter account** with 124 national and international followers and about 5-10 posts per month. HDS-LEE uses **advertising services**.

After the Helmholtz Virtual Data Science Career Day in September 2020, HDS-LEE launched a **job-advertising mailing list** (hds-lee-newsletter@fz-juelich.de). Eight people are subscribed to this mailing list, no posting so far.

Furthermore, HDS-LEE launched an **event-advertising mailing list** for scientist interested in talks and workshops organized by the school (hds-lee-event@fz-juelich.de). 22 subscriptions, one posting per month.

MARIO RÜTTGERS – HDS-LEE

What Hurricanes and Nasal Breathing Have in Common

Many patients who have nasal surgery do so in the hope of being able to breathe better. But the results are often disappointing due to a lack of knowledge about how air flows through the nose. At HDS-LEE, Mario Rüttgers is applying data science methods to reveal the small hurricanes that occur in the nose when we inhale. He wants to use these insights to help make future nasal surgeries more successful

On the wall of Mario Rüttgers' study at RWTH Aachen University hangs a medical illustration of the human nose with the corresponding Latin names: *cavum nasi, conchae nasales, septum nasi.* It is rather unusual for an engineer to have to learn medical terms like these. Rüttgers is a doctoral researcher in the first cohort of postgraduate students at HDS-LEE, the Helmholtz School for Data Science in Life, Earth and Energy. However, his main research interest is not anatomy but rather fluid mechanics – in conjunction with artificial intelligence methods. Together, these two areas have led him to a doctoral research topic where airflows play a key role – the nose.

Rüttgers first got in touch with the field of fluid mechanics while studying for his BA in engineering at RWTH Aachen. He was working on a simulation of casting processes, as computers can be used to calculate in advance how a mass will flow into a casting mold. The simulation sparked his interest: "Even then, I found it really fascinating that computers could be used to simulate how things behave in reality."

Korean typhoons as giant study objects

And the link between fluid mechanics and artificial intelligence methods then came about thanks to "a slight detour," as Rüttgers describes it. He ended up studying in Korea – a move that was motivated by love, as his then girlfriend and now wife is from South Korea. A scholarship from the South Korean government paved the way for the engineering graduate who exudes Far-Eastern calmness. He studied the language for three semesters so he could get to grips with his new day-to-day life in South Korea, and the MA he subsequently studied for at Pohang University of Science and Technology led him to data science methods.

The question that Rüttgers focused on during his time abroad is highly relevant to the coastal region near Pohang on the Sea of Japan: How do typhoons move, and how can we predict how severe they will be? The coasts of the Korean Peninsula and the neighboring countries of China, Taiwan, and Japan are severely affected by these Pacific hurricanes, which are becoming stronger and more destructive all the time due to global warming.

Rüttgers' task was to write an algorithm that learns from satellite data and flow simulation data while incorporating parameters such as the sea surface temperature, air pressure, and wind speed. Predictions are currently still being made using simulations that require huge processing times and expensive hardware. The learning involved in artificial intelligence methods, on the other hand, only requires significant effort on a once-off basis. The simulations based on these methods can then be created in real time – at low cost and without supercomputers. The incredible speed at which predictions can be made using the new algorithm is especially advantageous when it comes to issuing warnings at the right time and saving lives in the process.

In February 2019, Rüttgers and his wife returned to Germany with their daughter, who was born in South Korea. And he brought his master's degree and a wealth of insights into artificial intelligence methods with him. Back at home, this enabled him to get started in a new research field at the Institute of Aerodynamics at RWTH Aachen. Rüttgers now conducts bio-fluid mechanics research, which looks at flows in artificial organs and implants such as artificial heart valve prostheses and blood pumps, as well as flows in the human body in general, for example in the airways.

Flow simulations for a very small area

The flow simulations, which thanks to Rüttgers' efforts do not focus on the expanses of the Sea of Japan but rather on a significantly smaller area, namely the human nose, aim to gain a better insight into symptoms commonly seen in ENT medicine. Many people experience breathing problems due to constrictions and curvatures in the nasal cavity. Surgical intervention is used in severe cases with the goal of providing relief for the affected patients. However, studies have shown that, in many cases, an operation does not result in any significant improvement and can even lead to other health issues. This is because doctors are still using CAT scan images as their primary reference point for surgical alterations in the nasal cavity.

Rüttgers is using his knowledge of artificial intelligence to contribute to the research goal of using a differentiated flow simulation to support doctors in making their diagnoses and, if surgery is being considered, to provide information that can help ensure its success. The ultimate aim is to make it easy for doctors to see how an operation could impact breathing using a computer.

The simulation calculates various physical parameters, such as the pressure at which air flows into the nose: How much energy does the patient need to exert in order to breathe, both before and after the operation? The researchers also consider the friction that inhaled air exerts on mucous membranes in the nose. This can increase if the patient finds it easier to inhale after an operation. However, this can also cause inflammation over time, an undesirable side effect of an apparently successful operation. Another key aspect of the breathing process is that air needs to be sufficiently warmed up so that it is not too cold when it reaches the lungs, creating favorable conditions for inflammation and infection. This means it has to make certain "detours". To do so, it circulates in what are called the nasal conchae, where it is warmed up by the nasal membranes before flowing onward to the lungs. If these pathways are no longer in place after an operation, health problems can occur in the long term.

Planning operations better with artificial intelligence

Rüttgers' job is to build an automatic "data pipeline" with the help of AI methods. In this pipeline, pathologies are automatically detected and localized on the basis of CAT images, for example, or the surface of the airways is extracted and prepared for simulation. As with a large-scale typhoon simulation, the area inside the nose is then divided into billions of small elements, and various parameters such as pressure, friction, and temperature can be calculated for these elements. Various AI elements need to work together to obtain a result at the end of these complex computing operations. This requires a lot of processing power, which is why Rüttgers also works at the Jülich Supercomputing Centre (JSC), alongside his position at RWTH. The JSC gives him access to hardware he can use for his calculations. Thanks to Rüttgers' data pipeline, setting up the simulations is not only more cost-effective but quicker as well – making them a viable option for everyday clinical applications. The other members on the research project then use Rüttgers' AI methods as the basis for building the software for virtual operations and the corresponding visualization system, so doctors can use the new digital tool with ease.

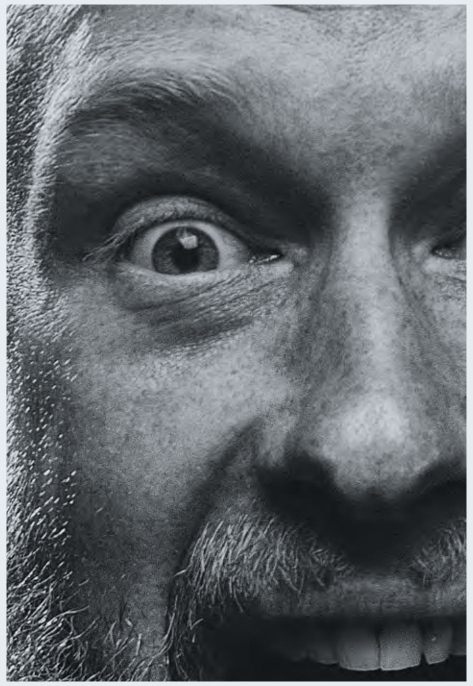
"The added value of our method is that you can use it to simulate an operation on a computer first and then see what effects it might have," Rüttgers explains. This makes it easier to assess any complications that might occur in each region of the nose, and an alternative surgical approach can be reviewed in advance in the same way. This provides surgeons with a precise recommendation as to the best course of action –"and it's all completely non-invasive," Rüttgers emphasizes. Applying a method like this would mean significant relief for patients and greater certainty for surgeons that they are choosing the right approach. Moreover, the health system could save costs by boosting the efficiency of treatment.

This combination of artificial intelligence and biomedical research makes Rüttgers and his project an excellent fit for the Life Science division at HDS-LEE. "As a postgrad associated with HDS-LEE, I benefit first and foremost from the training opportunities such as soft skills courses and of course the chance to network. Talking to other people who come across similar problems or challenges in their research is extremely valuable." In addition to the biosciences and medicine, doctoral researchers at HDS-LEE focus on energy systems and materials as well as the geosciences. In other words, they work on a broad spectrum of applications for data science methods that aim to arrive at a better understanding of complex systems – ranging from vast to minuscule. And this is another reason why Rüttgers is in the right place at HDS-LEE – with his unique research background ranging from typhoons to nasal breathing.

Author: Constanze Fröhlich

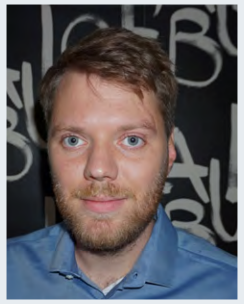
https://www.helmholtz-hida.de/en/activities/news/newsdetail/what-hurricanes-and-nasal-breathing-have-in-common/

Credits: Mario Rüttgers: private/Nose in focus: Alexander Krivitsk/Unsplash



CHRISTIAN GERLOFF – HDS-LEE

Where the Neurons Are Firing



Christian Gerloff completed his degree in electrical engineering. Now, he's tracking down signals from one of the most mysterious and complex systems in existence. As a doctoral researcher at the HDS-LEE graduate school, Gerloff conducts research that unravels more hidden secrets of the human brain – thanks to Data Science methods

Before Christian Gerloff even had the chance to experience an aha moment as he sat at his computer, his work started with people – at the University Hospital Aachen, for example – wearing strange, black hoods over their heads. Various cables, octodes, and sensors were attached to the hoods, and infrared light was directed through the test subjects' scalps. The reflected light was measured in turn; specifically, while the test subjects were in varying states – such as playing computer games in pairs or simply relaxing.

By setting up these experiments, neuroscientists at the hospital want to use functional near-infrared spectroscopy (fNIRS) to determine which regions in the subjects' brains are active in the respective states. This can be detected by the change in oxygen levels in the blood, which can be measured using the infrared light. When a region of the brain is active, the flow of blood to this area typically increases. The oxygen level in this location also rises as a result, and it can be presumed that a particularly high number of neurons fire in response. Researchers want to use the data from these measurements to draw conclusions.

And this, to put it briefly, is the point where Gerloff's work starts. Gerloff is a doctoral researcher, and his thesis titled "Machine learning and Bayesian models in the neurosciences" aims at reliable insights into how the human brain is organized – with respect to basic research as well as subsequent therapeutic or clinical applications.

Enigmatic brain

"The brain," Gerloff says, "is one of the most complex systems in existence." But for scientists, it remains a black box in many ways. That's why the young researcher is happy when he is able move a step forward in deciphering the links between individual areas of the human brain, or how different individuals' brains behave in relation to one another as they interact.

Ultimately, these findings could be used to explore the correlation between social experiences and biological systems, such as whether there are characteristic neuronal states in which a mother and her child can respond to each other more effectively in emotional terms. Gerloff's research is also a starting point for fields of clinical application, such as achieving a better understanding of the neurological conditions behind diseases such as ADHS or predicting cognitive impairments that strokes will cause in individual cases.

Advancing the neurosciences as an engineer

But Gerloff is not a neuroscientist – he is an engineer. He earned his master's degree in electrical engineering from RWTH Aachen University. His thesis looked at how sensor data from modern vehicles can be used to predict when and what types of defects will occur in vehicle parts. During this time, Gerloff also worked on topics in various disciplines using machine and deep learning methods – in other words, with artificial neuronal networks that imitate the way the brain functions. This is how he came into contact with the neurosciences; the field captured his interest, and he decided to dig deeper.

And this is exactly what made him a suitable candidate for the HDS-LEE graduate school in Jülich. The School for Data Science in Life, Earth and Energy was founded by Helmholtz in March 2019 and represents a partnership between RWTH, the University of Cologne, the German Aerospace Center, the Max Planck Institute for Iron Research, and Forschungszentrum Jülich.

HDS-LEE aims to encourage a disciplinary approach between the data sciences and natural sciences. The school has adopted this focus in response to the ever-increasing and increasingly complex data generated in most of the natural sciences. This requires experts who understand the questions addressed by the natural sciences and have the expertise that is needed to obtain the greatest possible insights from such data. To this end, HDS-LEE supports young, talented scientists in the fields of mathematics as well as the computer, natural, and engineering sciences who are driving the development of data science methods.

Gerloff has been part of the first cohort at HDS-LEE since mid-2019. Roughly speaking, the approach behind his doctorate is to analyze the brain as a functional network – and to use data science methods to investigate exactly how it operates.

To do this, Gerloff applies methods from the field of machine learning – including algorithms and statistics as well as Bayesian models, which calculate the degree of probability – to process data from experiments or databases that are obtained, for example, using functional near-infrared spectroscopy (fNIRS) or functional magnetic resonance imaging (fMRI). A large part of his work consists of programming high-performance computers, as well as testing his implementations and methods. In addition, Gerloff continually works to improve his methodological understanding and refine the methods he uses to ensure that his studies will result in robust research findings.

The brain as source of inspiration and field of application

The young researcher is fascinated by his field and says, "The brain's functions and its neuronal networks determine our cognitive capabilities. In other words, they define us to a certain extent. At the same time, they are the starting point for a wide range of methods in data sciences." After all, these methods are inspired in part by the way the human brain processes information. "But at the same time, the brain and the data we now have on the brain are so complex that we need to improve our understanding of the methods and continually enhance modern methods so we can achieve better insights into the brain," says Gerloff. "In this sense, the brain can serve as a source of inspiration and a field of application at the same time." Which is why combining brain research and data science is a wonderful symbiosis.

"Exchange" is the magic word here. And this is exactly what HDS-LEE encourages with its interdisciplinary support. The professors supervising Gerloff's thesis are Professor Kerstin Konrad, who heads the teaching and research area of childhood and adolescent clinical neuropsychology at University Hospital Aachen as well as an institute at Forschungszentrum Jülich; Professor Danilo Bzdok from McGill University and the Mila Quebec Artificial Intelligence Institute; and Raul Fidel Tempone, a professor of mathematics at RWTH Aachen.

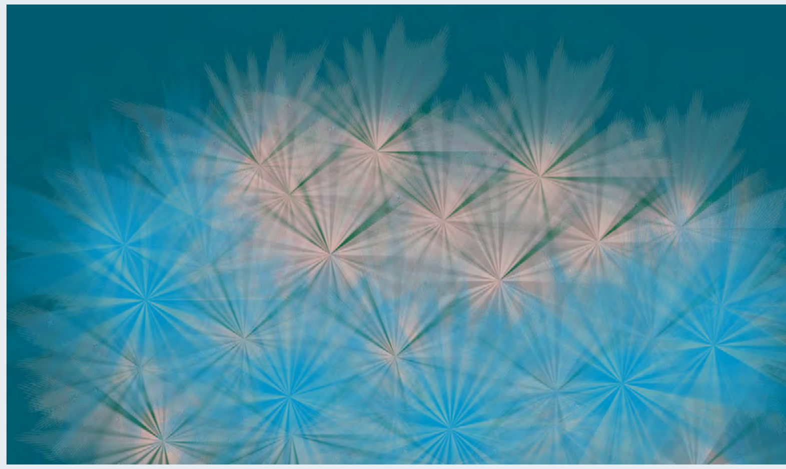
"HDS-LEE also makes it possible to exchange ideas and information beyond our respective research fields," Gerloff says – and not just with his supervisors but also with doctoral researchers in other disciplines. "This gives us an insight into the other faculties and can also inspire a better understanding of the problems in our own areas or different ways of approaching them." Different perspectives are key, and both sides benefit as a result. Doctoral researchers at HDS-LEE attend specialist lectures, seminars, and conferences together. "This means you're not just working for yourself in your own little bubble," says Gerloff. "Instead, you discuss both professional and personal matters with the other doctoral researchers."

If you ask Gerloff whether he has had any aha moments in his research yet, he grins and says: "Yes, and the exact opposite, too!" He says that he's sometimes been very happy when he has gained new insights into neuronal interconnections, but that he has sometimes also realized later on that his anticipation was premature, as some clue or another ultimately came to nothing. That's the way research goes. You need to be willing to sit tight and have a certain tolerance for frustration. Gerloff appears to have these qualities. He is the type of person who thinks before he answers and, above all, doesn't get ahead of himself. Gerloff has a warm personality and likes the fact that his research focuses on people first and foremost.

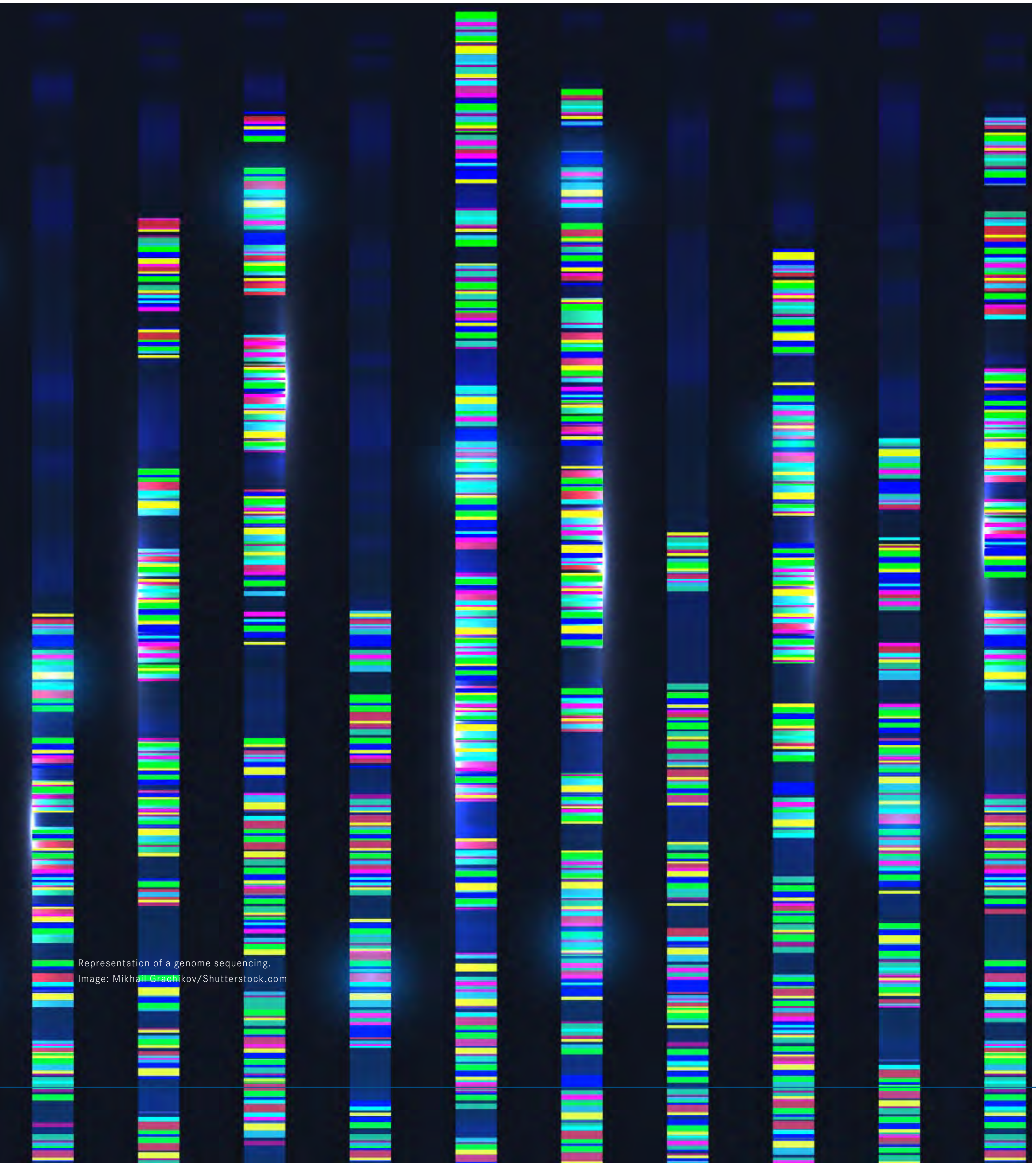
Author: Andrea Walter

https://www.helmholtz-hida.de/en/activities/news/newsdetail/where-the-neurons-are-firing/

Credits: Christian Gerloff: private/ Artistic data visualization based on the data of C. Gerloff: Michael Schmitz for Helmholtz/HIDA



MUDS MUNICH SCHOOL FOR DATA SCIENCE HELMHOLTZ | TUM | LMU



Representation of a genome sequencing.
Image: Mikhail Grachikov/Shutterstock.com

MUNICH SCHOOL FOR DATA SCIENCE @ HELMHOLTZ, TUM & LMU (MUDS)

The goal of the MUDS is to educate the next generation of Data Scientists at the interface of Data Science and four different application areas with a structured doctoral program: biomedicine, plasma physics, earth observation and robotics. Although the specific research questions in these four domains are different, all of them have both purely data- and model-based research approaches. MUDS researchers aim to explore new ways to connect these two poles.

Research areas:
Biomedicine, plasma physics, earth observation and robotics.

Partners of MUDS:
Partner of MUDS: Helmholtz Zentrum Munich, German Aerospace Center, Technische Universität Munich, Ludwig-Maximilians-Universität Munich as well as the Leibniz Computing Center and the Max Planck Computing & Data Facility

lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften

IPP Max-Planck-Institut für Plasmaphysik

MPCDF Max Planck Computing & Data Facility Rechenzentrum Garching der Max-Planck-Gesellschaft

HelmholtzZentrum münchen Deutsches Forschungszentrum für Gesundheit und Umwelt

LMU Ludwig-Maximilians-Universität München

TUM Technische Universität München

DLR

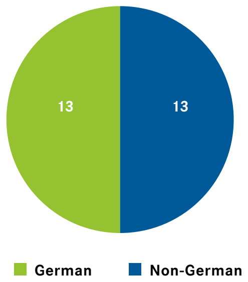


Fig. 21: Recruitment of doctoral researchers from abroad at MUDS 2019–2020 (excl. associated doctoral researchers).

Applicant situation and recruitment

MUDS filled 17 PhD positions in 2019, and 9 in the following year 2020 (associated doctoral researchers not included). 129 candidates from 36 nations applied for the first round in 2019; in 2020, even 313 applications were submitted from 51 countries.

The 48 PhD researchers at MUDS (incl. associated doctoral researchers) come from a total of 18 different countries; among them are 29 Germans. One third of the PhD researchers are women (15). 22 additional researchers, among them 9 women, are associated with MUDS as doctoral researchers; they all come from Helmholtz Zentrum München (HMGU).

Events and networking

In 2020, the MUDS has organized eight events:

4 Networking events

- 10.02 - 11.02.2020, Interview days including poster session and dinner, Technical University Munich, Garching, 2-day MUDS recruiting event (76 participants, 20 female, 56 male, 31 MUDS, 42 Helmholtz, 34 outside Helmholtz)
- 21.07.-22.07.2020, Virtual recruiting event (MUDS Health track), 2-days virtual MUDS recruiting event (MUDS Health Track), (46 participants, 18 female, 28 male, 8 MUDS, 17 Helmholtz, 29 outside Helmholtz)

- 06.11.2020 (2.5 h), MUDS Welcome Event, virtual, Welcome, Introduction to PhD administrative matters, Q&A, teambuilding game, (27 participants, 10 female, 17 male, 27 MUDS, 26 Helmholtz, 1 outside Helmholtz)
- 09.12.2020 (2h), MUDS PI Workshop, virtual, connect MUDS core PIs with interested partners and with each other, to create synergies and to initiate partnerships for MUDS proposals in the ongoing call, (23 participants, 7 female, 16 male, 11 MUDS, 17 Helmholtz, 6 outside Helmholtz)

3 Courses

- 02.07-03.07.2020 (each half day), Course on Good Scientific Practice: online course to inform about the rules of research integrity and good scientific practices and how to avoid research misconduct, (18 participants, 7 female, 11 male, 18 MUDS, 16 Helmholtz, 2 outside Helmholtz)
- 21.08.2020-18.09.2020 (each Friday), Peer-to-Peer Course "Data Science for Researchers Practical Training", online course organized and taught by fellow MUDS students on Fundamentals & Best-Practices for Software development, (15 participants, 4 female, 11 male, 15 MUDS, 12 Helmholtz, 3 outside Helmholtz)
- 11.11.-10.12.2020 (18 x half day + exam day), Data Science Block Course, virtual, 2 week online block course covering several data science-related topics, (17 participants, 6 female, 11 male, 17 MUDS, 15 Helmholtz, 2 outside Helmholtz)

Ongoing lecture

- 05.08.20 – ongoing, MUDS Seminar Series, student progress reports, on average biweekly, virtual, talks of guest speakers invited by students and progress reports by MUDS doctoral researchers or talks by MUDS PIs

Publications

A total of ten first author publications (journal/conference) were published and four co-authored publications (journal/conference).

First-author publications:

1. Boushehri, S. S., Qasim, A. B., Waibel, D., Schmich, F., & Marr, C. (2020), Annotation-Efficient Classification Combining Active Learning, Pre-Training and Semi-Supervised Learning for Biomedical Images, bioRxiv.
2. Dorigatti, E. and Schubert, B., Graph-Theoretical Formulation of the Generalized Epitope-Based Vaccine Design Problem, PLoS computational biology 16.10 (2020): e1008237.
3. Dorigatti, E. and Schubert, B., Joint Epitope Selection and Spacer Design for String-of-Beads Vaccines, Bioinformatics 36.Supplement_2 (2020): i643-i650.
4. Feng, J., Durner, M., Marton, Z. C., Balint-Benczedi, F., & Triebel, R. (2019), Introspective Robot Perception Using Smoothed Predictions From Bayesian Neural Networks.
5. Geisler, S., Zügner, D., & Günnemann, S. (2020), Reliable Graph Neural Networks via Robust Aggregation. Advances in Neural Information Processing Systems, 33.
6. Höllbacher, B., Balázs, K., Heinig, M., & Uhlenhaut, N. H. (2020), Seq-Ing Answers: Current Data Integration Approaches to Uncover Mechanisms of Transcrip-Tional Regulation, Computational and Structural Biotechnology Journal.
7. Kondmann, L., & Zhu, X. X. (2020), Measuring Changes in Poverty with Deep Learning and Satellite Imagery.
8. Kondmann, L., Haeberle, M., & Zhu, X. X. (2020), Combining Twitter and Earth Observation Data for Local Poverty Mapping, in: NeuRIPS Machine Learning for the Developing World Workshop (pp. 1-5).
9. Loureiro, H., Becker, T., Bauer-Mehren, A., Ahmidi, N., & Weberpals, J. (2020, October), Improving Predictive Ability of Survival Models: Comparison of Multiple State of the Art Models, in Pharmaco-Epidemiology and Drug Safety (Vol. 29, pp. 35-36).
10. Rath, K., Albert, C. G., Bischl, B., & von Toussaint, U. (2020), Symplectic Gaussian Process Regression of Hamiltonian Flow Maps. arXiv preprint arXiv:2009.05569.

Co-authored publications:

1. Waibel, D. J. E., Boushehri, S. S., & Marr, C. (2020), InstantDL-An Easy-to-Use Deep Learning Pipeline for Image Segmentation and Classification, bioRxiv.
2. Brunner, A. D., Thielert, M., Vasilopoulou, C., Ammar, C., Coscia, F., Mund, A., Richter, S., & Mann, M. (2020), Ultra-High Sensitivity Mass Spectrometry Quantifies Single-Cell Proteome Changes Upon per-Turbation, bioRxiv.
3. Lee, J., Humt, M., Feng, J., & Triebel, R. (2020), Estimating Model Uncertainty of Neural Networks in Sparse Information Form, in: International Conference on Machine Learning (pp. 5702-5713). PMLR.
4. Eisenberger, M., Toker, A., Leal-Taixé, L., & Cremers, D. (2020), Deep Shells: Unsupervised Shape Correspondence with Optimal Transport, arXiv preprint arXiv:2010.15261.

Further news from MUDS

3 Doctoral researchers are funded by **industry partners** (2 by Roche Diagnostics GmbH, 1 by Boehringer Ingelheim Pharma GmbH), 23 Doctoral researchers are funded by other third party funding.

Communication/Marketing

MUDS has a **website** (www.mu-ds.de), a **Twitter account** with 595 followers (Jan 2021), and a **mailing list**. MUDS uses **advertising services**.

KARIN HROVATIN – MUDS

Hope for Diabetics

Diabetes, a widespread disease, is still incurable in most cases. But with the help of data science, this could soon change. At MUDS, doctoral researcher Karin Hrovatin is researching the properties of the many cells involved in insulin production. In the future, her findings should help to stimulate defective cells in the pancreas to produce the vital hormone again

As a child in Slovenia, Karin Hrovatin loved drawing and math. She was particularly fond of a primary school exercise book called "Math is a Game" that required coloring in pictures to solve math problems – so fond of it that once she finished the exercises in her copy she started again in her classmate's workbook.

Today, Hrovatin is a PhD candidate at the Munich School for Data Science (MUDS). But she's still using color, drawing and visualization to solve big problems. "Most people see data science as computer programming, statistics, and math," Hrovatin says. "But actually, drawing and plotting visualizations is one of the key aspects of data science."

Using scatter plots that resemble rainbow-colored clouds or abstract art, Hrovatin is taking on a problem that affects nearly 1 in 10 adults worldwide: Diabetes, a chronic condition which can lead to amputations, blindness, heart attacks and strokes. "Diabetes is a very problematic disease," says Hrovatin. "And it's very widespread, which makes it a hot topic to try to solve."

Better understanding the body's insulin factory

After earning a degree in biotechnology at the University of Ljubljana and studying bioinformatics and diabetes at the University of Edinburgh, Hrovatin realized that she was more interested in numbers than hands-on lab work. But the Slovenian data scientist wanted to keep working on diabetes, one of the most pressing problems in public health today.

A unique interdisciplinary data science program run by the Helmholtz Association enabled her to do both, by applying cutting-edge data science methods to better understand the progression of diabetes at the cellular level. Now a doctoral researcher at the Helmholtz Information and Data Science Academy's Munich School for Data Science, or MUDS, Hrovatin is working on ways to better understand beta cells, the body's insulin factories. Located in the pancreas, beta cells respond to rising levels of glucose, also known as blood sugar, in the bloodstream by producing the hormone insulin. That, in turn, signals muscle cells to absorb and store blood sugar to use later.

But when the beta cells break down, they stop producing enough insulin – and the body stops absorbing blood sugar. That, in turn, causes type 2 diabetes, one of the most widespread and fast-growing non-communicable diseases in the world today.

To find a cure, researchers are working to better understand why beta cells stop working. The old assumption was that eating too much sugary food forced the beta cells to work too hard, eventually causing them to wear out. But researchers now know that beta cells aren't all the same. They vary from person to person, or even between neighboring cells. Understanding what makes beta cells tick – and, sometimes, stop ticking – is at the heart of Hrovatin's work.

While researchers once assumed that worn-out beta cells were beyond repair, new discoveries have shown that some beta cells can be revived – but it's not yet clear which ones respond to treatment. "Beta cells change during aging, or from stress. We're looking for the differences between healthy and diseased beta cells, so we can figure out how to regenerate cells or bring back their function," she says. "If you could convert them back to their healthy state, you could restore beta cell function to the pancreas."

An "atlas" of insulin-producing beta cells

The lab she's part of at the Helmholtz Center Munich's Institute for Computational Biology is focused on "single-cell sequencing" data analysis a technique that enables biologists to directly examine cell traits in diabetic mice. Despite its huge potential, the research is conducted on a microscopic scale: Hrovatin's collaborators extract pancreas cells from the rodents and split the organ tissue up into single cells, to study their properties and how they respond to treatments and stress.

The beta cells are lined up in a tiny tube and encapsulated in oil droplets, then individually "tagged" with a unique identifier. A given data set might be based on just 10,000 cells, an amount barely visible to the naked eye.

Hrovatin's ultimate goal is to combine data from many different cell types and experiments to create a sort of "atlas" of beta cells, understanding which ones share properties and why their metabolic function differs. That could guide future researchers to better, more personalized treatments for type 2 diabetes. The Helmholtz Diabetes Center is at the forefront of the research. "There's a lot of data collected already," Hrovatin says, "which makes diabetes a good problem for data analysis."

Using machine learning and data science, Hrovatin plans to analyze what cell traits different disease types and treatments have in common – creating a sort of atlas of cell subtypes. "I could use that information to predict how cells will likely respond to treatment in a living organism," Hrovatin says, "based on how they respond in cell culture." That, Hrovatin hopes, will be an important first step towards therapies for people with diabetes.

Important step: making data sets comparable

Before she can make meaningful comparisons of cell types and function, though, she has to make sure data sets match up. In an ideal world, all researchers would use the same methods in their experiments, making it possible to easily compare results to see how cells react to different stresses and treatments.

But this is not an ideal world. When she arrived in Munich this summer, she realized the scale of the problem: "There are different mouse models, different disease types, different lab protocols," Hrovatin says. "There's a lot of variation, and that's a big challenge."

As a result, she's spent the first four months of her PhD finding ways to make data sets line up, to make sure cells can be compared in a meaningful way. "It's important to bring the data sets together, but first you have to make sure you're analyzing biological effects, not technical effects," Hrovatin says. "I'm hoping to collaborate with biologists to check the results." So far, she's been able to draw on the expertise of diabetes and bioinformatics researchers at the Helmholtz Center Munich.

Interdisciplinary suggestions by the MUDS

The Alpine backdrop of Munich reminds the Slovenian researcher of her native Ljubljana, the capital of Slovenia. She's enjoyed exploring the city by bicycle – and commuting to her lab, despite the strange and isolating conditions of the coronavirus pandemic. MUDS, meanwhile, provides a growing network of other Doctoral researchers with similar interests.

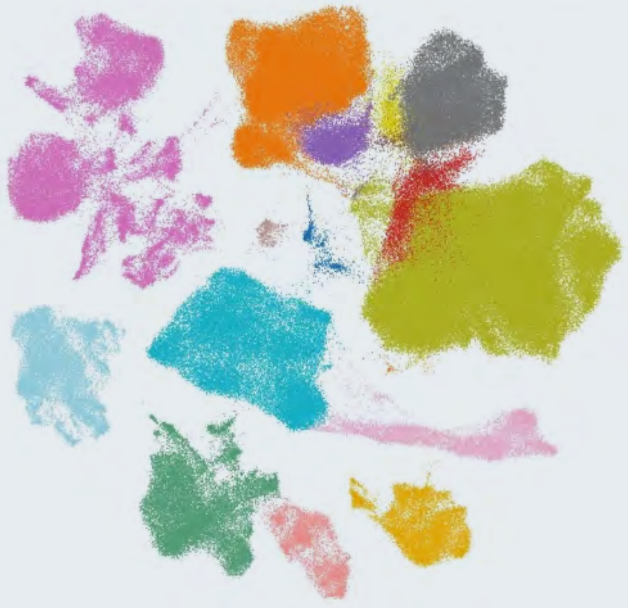
Part of the Helmholtz Information and Data Science Academy, MUDS connects the Helmholtz Center Munich with the Max Planck Institute for Plasma Physics and the German Aerospace Center. Her peers include experts in robotics, plasma physics and biomedicine – all united by a shared focus on applying data science in new ways. Seminars bring the different doctoral researchers together regularly online, until they can start meeting in person. Says Hrovatin: "It's great – I can learn from people in other fields and get ideas to apply to biology as well."

Hrovatin still finds time to indulge her childhood passion for drawing. In her free time she sketches fanciful fashion and haute couture, for example. But her energy these days is poured into a different kind of art: Painting colorful visualizations with data points and cell types, hoping to one day heal people with diabetes.

Author: Andrew Curry

https://www.helmholtz-hida.de/en/activities/news/newsdetail/datenbilder-aus-der-insulinfabrik/

Credits: Karin Hrovatin: private/Graphical scheme of the distribution of 16 different cell types found in the pancreas: Karin Hrovatin



LAURA MARTENS – MUDS

Healing by Gene Therapy?

Many diseases such as cancer or rheumatism are based on genetic defects. Will it be possible to treat them with gene therapy in the future? To get closer to this goal, MUDS doctoral researcher Laura Martens is trying to decipher the gene regulatory cell codes. She is being helped by data from the international megaproject Human Cell Atlas

In October 2016, leading scientists from around the world met in London to discuss setting up a Human Cell Atlas. It would be a once-in-a-century project – because creating an overview of the human cell inventory would make it possible to define the basic cellular principles of health and describe a wide range of diseases with greater precision. Even though researchers have long been analyzing cells as the most fundamental building blocks of life, we still know very little about them. At the same time, we need a precise understanding of the various cell types, as this could enable new insights into the development and treatment of very different diseases – from autoimmune diseases like rheumatism, through cardiovascular diseases and chronic inflammation processes, to cancer.

Since October 2020, Laura Martens, who is based in Munich, has been involved in efforts to shed some light on the secrets of the human cell and is working toward this goal with data resources from the Human Cell Atlas. The young physicist from Bremen just completed a master's degree in computational biology in the UK last year and was looking to return to Germany for her doctorate. That's when a tweet from Fabian Theis at Helmholtz Zentrum München (HZM) came up on her Twitter account. He was writing to let her know about the new round of calls for applications at MUDS, the Munich School of Data Science. The school is part of the Helmholtz Information & Data Science Academy (HIDA), Germany's largest postgraduate training network in the data sciences. Her application was successful, and Martens is now conducting research at MUDS on her PhD project, which is titled "Unraveling the gene regulatory code using single-cell multi-omic data." "Data science with biology or biomedicine – that's exactly what I want to be doing!" Martens' work is supervised by two principle investigators, both of whom are experts in the field of computational biology. She refers to them as "Fabian and Julien": Fabian Theis, Head of the Institute for Computational Biology at HZM and Julien Gagneur, Professor of Computational Molecular Medicine in the Department of Informatics at the Technical University of Munich.

Using the data of a megaproject

Martens isn't alone in her enthusiasm for biomedical topics – people at over 1,200 research institutions on every continent are working to complete the Human Cell Atlas (HCA). Their aim is to provide an even better description of the human body. Eighty-two institutes are involved in the project in Germany alone, including four Helmholtz Centers (Helmholtz Zentrum München, German Cancer Research Center, German Center for Neurodegenerative Diseases, and the Max Delbrück Center for Molecular Medicine). The Atlas serves as a repository in which data from various cell analyses are made available to researchers for further study. This makes the HCA an outstanding example of how open science can work in practice – a topic that Helmholtz is also actively advancing with the goal of achieving benefits for science and society. As a large-scale, global research project, the HCA was made possible by a technology called single-cell RNA sequencing that was developed only 15 years or so ago.The technology can be used to analyze which RNA molecules are present in an individual cell, allowing conclusions to be drawn about the specific function – or dysfunction – of that cell.

Before single-cell sequencing was an option, the only analyses researchers could carry out on the genetic material in a tissue sample were unspecific, and they were unable to look at the differences between the cells – even though there are many very different types of cells in an organism. Martens, whose project draws on data that were acquired using this process, has an appealing way of describing it: A tissue sample of this kind can be imagined as a sort of "smoothie"; that is, a nice mix of various types of cells. When you analyze this material, she says, you therefore get an average of all the cells contained in the mix. "Now, with single-cell," Martens continues, "you have the smoothie, but you can tell that it contains three strawberries and five blueberries; so now, you can look at individual cells to see what's happening inside them."

Not all cells are equal

The actual object of her research – the individual cells, or, to use the same metaphor, the blueberries – is extremely complex. This is because all cells are not the same. While every cell in the human body contains the identical copy of the genome, around three billion base pairs, they exhibit dramatic variations in their morphology and behavior patterns. But even cells that have been assigned to the same type up until now are different from one another. For a long time, it was thought that there are around 300 different types of cells in the human body, but thanks to a much more detailed knowledge of their biochemical "appearance," we now know that this figure must be significantly larger. This is because single-cell RNA sequencing makes it possible to differentiate many types of cells in further sub-categories. This can be seen in the RNA, which transmits the blueprints for proteins that are encoded in the DNA to the ribosomes, the "workshops" of the cells, where the corresponding proteins are synthesized. As a result, looking at a cell's complete RNA catalog makes it possible to see which genes are active there and how the cell is regulated.

The fact that a cell becomes a very specific type of cell can be attributed to various mechanisms known as gene regulatory elements, which have yet to be examined in detail. These elements determine how a gene of the cell DNA is "expressed" – for example, whether and how the transcription of a gene or the synthesis and degradation of RNA takes place. These various steps within the biosynthesis of proteins are a crucial factor in how the genetic information in a cell manifests itself. They indicate the function of the cell and how it develops. In other words, why do the blueberries become blueberries?

Martens now wants to look at some of these gene-regulating mechanisms – and decode their interactions. She started with a process that comes at the very end of protein biosynthesis, the degradation of RNA. "Of course I'd like to understand everything at once," says Martens, "but I'm starting by having a try with this final element that involves the degradation of RNA." This particular biomolecule, which carries the genetic information of the DNA and is needed to initiate the actual synthesis of proteins, only ever exists for a relatively short period. But if the RNA lasts for longer, for hours or days, for instance, more proteins can be produced; the RNA is used multiple times. In other words, the speed of RNA degradation has an effect on both the gene expression and the many other gene regulating mechanisms – and, like these, is encoded in the DNA sequence.

Unraveling the enigmatic combinatorics of gene regulation with Data Science

Martens knows that the interconnections between various types of genetic information are complex, and she wants to shed further light on these connections as she pursues her project. "You have to imagine it as an enormous network or enormous combinatorics that is interconnected, but we still don't even know exactly how it actually works. It's incredibly complex, and these various elements can't even be properly unraveled yet." She adds: "I feel overwhelmed by the complexity sometimes, too!" But it's evident that the researcher doesn't see this as reason for giving up; on the contrary, it's what really motivates her. Because a moment later, she's bubbling with enthusiasm for her new research terrain. "That's what's so incredibly fascinating about it at the same time; that this network has developed like this through evolution! It's so complicated that you really can't get to the bottom of it just by looking at the data. And that's also why deep learning is coming into play now."
This is because wherever data becomes unmanageable for people, artificial intelligence can help. Martens is using the complex data from single-cell sequencing as the basis for developing a general, computer-assisted framework – in other words, she is writing a program. This program is designed to facilitate systematic analyses that deal with the respective gene-regulating mechanisms specific to the cells. Martens' goal is to make it flexible enough so that it can be generalized at the level of various single-cell sequencing data and is capable of modeling multiple steps involved in gene expression, from the beginning of transcription up to the breakdown of the RNA. In order to make it possible to read the gene regulatory code of cells in the future, Martens also wants to use machine learning models which are intended to help interpret the data that are input in the program. "The single-cell data reveal which genes are expressed and how pronounced this expression is. But our goal is to understand why this is the case. So I want to know which input was especially important in a dataset." A program of this type could, for example, be useful for researchers who want to understand the gene-regulating mechanisms of certain types of cells that are important for the functioning of an organ. They include the scientists in Fabian Theis' research group, who are working on the Human Lung Cell Atlas, another project relating to the Human Cell Atlas.

"We don't even know what's broken in many cases"

During the first few months of her PhD, Martens' main focus was getting the data into a format that would make it possible to apply machine learning tools. The doctoral researcher's main job involves a lot of sitting in front of a computer. And even more so now due to the lockdown in Germany, which means that not only do her day-to-day interactions with the two labs take place online – her regular communication with the other new MUDS doctoral researchers across all the various disciplines does too. These contacts are something that Martens really appreciates about the program. In other words, the open-minded doctoral researcher doesn't live up to the cliché of the introspective IT nerd at all even though, as she says with a laugh, she spends the entire day fiddling with code. "When I started studying, I never would have thought that I would end up spending all my time writing program code," she admits.

But as someone who describes herself as more of a "theoretician," programming doesn't appear to be a huge obstacle, but rather her actual calling. Her knowledge of data science is helping Martens to understand at least a small part of this large, complex network that is gene regulation – and to support medical research in its pursuit of new therapy options in the process. This is what motivates Martens: “Many diseases go back to DNA. Once you understand how their building blocks interact with and influence each other, you can naturally try to repair things in specific places. Right now, we don’t even know what’s broken in many cases.” The prospect of gene therapy treatment, for instance using genome surgery, is a distant prospect on the horizon of the Human Cell Atlas. Laura Martens’ contribution to the unraveling of the gene regulatory codes of the cells appears to be akin to the way the initiators of the Human Cell Atlas described their historic research project: “ambitious, but achievable.”

Author: Constanze Fröhlich

https://www.helmholtz-hida.de/en/activities/news/newsdetail/zellcodes-entschluesseln-fuer-zukunftstherapien/

Credits: Laura Martens: private/ Representation of a DNA sequence: Mikhael Grachikov/ Shutterstock.com

11. OUTLOOK

In 2021, HIDA will continue to drive forward the many new activities and successes from 2020, as well as launching further new initiatives to provide additional training and networking opportunities in the field of Information & Data Science for all Helmholtz Centers and research fields. The goal will be to further intensify cooperation between researchers, to create spaces for knowledge exchange and training, and to actively support the Helmholtz Centers in recruiting new data science talent. Therefore, HIDA will substantially enlarge its talent recruitment initiatives e.g. by setting up an international employer branding campaign in the field of Information & Data Science to attract top-talent from leading research institutions worldwide. HIDA aims to maximize the perception of the Helmholtz Association as a key player in the field of Information & Data Science. Another focal point of HIDA will be further activities to promote all Helmholtz Centers and Helmholtz programs in the field of Information & Data Science.

It is HIDA’s major priority to continue to expand the HIDA Trainee Network, set up additional national and international exchange programs and partnerships and offer a variety of courses on Data Science methods. HIDA also plans to offer and to expand networking events (e.g. Virtual Career Day Vol. II; Virtual Datathon Challenge Vol. III, Data Scientist Award) and to take active part in and design recruiting events for the entire Helmholtz Association. Substantial marketing and communication activities complement these activities.

12. RESPONSIBILITY FOR REPORT / HIDA MAIN CONTACT

This report is checked and approved by the members of the Steering Committee of the Helmholtz Information & Data Science Academy (HIDA-Steer) in February 2021.

HIDA Main Contacts:

Susan Trinitz	Xenia von Polier
Adviser Strategic Initiatives, HIDA	*Press Officer, HIDA*

Helmholtz Information & Data Science Academy
Friedrichstraße 171, 10117 Berlin

IMPRINT

Published by:
Hermann von Helmholtz-Gemeinschaft
Deutscher Forschungszentren e.V.

Head Office and Bonn Office
Ahrstrasse 45, 53175 Bonn
Phone: +49 (0) 228 30818-0
Email: info(at)helmholtz.de

Strategic Initiatives Helmholtz Association
Helmholtz Information & Data Science
Academy (HIDA)
Friedrichstraße 171
10117 Berlin
Phone: +49 30 5498226-0
Email: hida(at)helmholtz.de

V.i.S.d.P. (Person responsible according to the German Press Law): Franziska Broer

Project Managers: Susan Trinitz,
Xenia von Polier

Editors: Susan Trinitz, Xenia von Polier,
Constanze Fröhlich (Freelance)

Design: Kai Kullen

Printing: Druck- und Verlagshaus
Zarbock GmbH & Co. KG
Frankfurt am Main

Correct as of March 2021