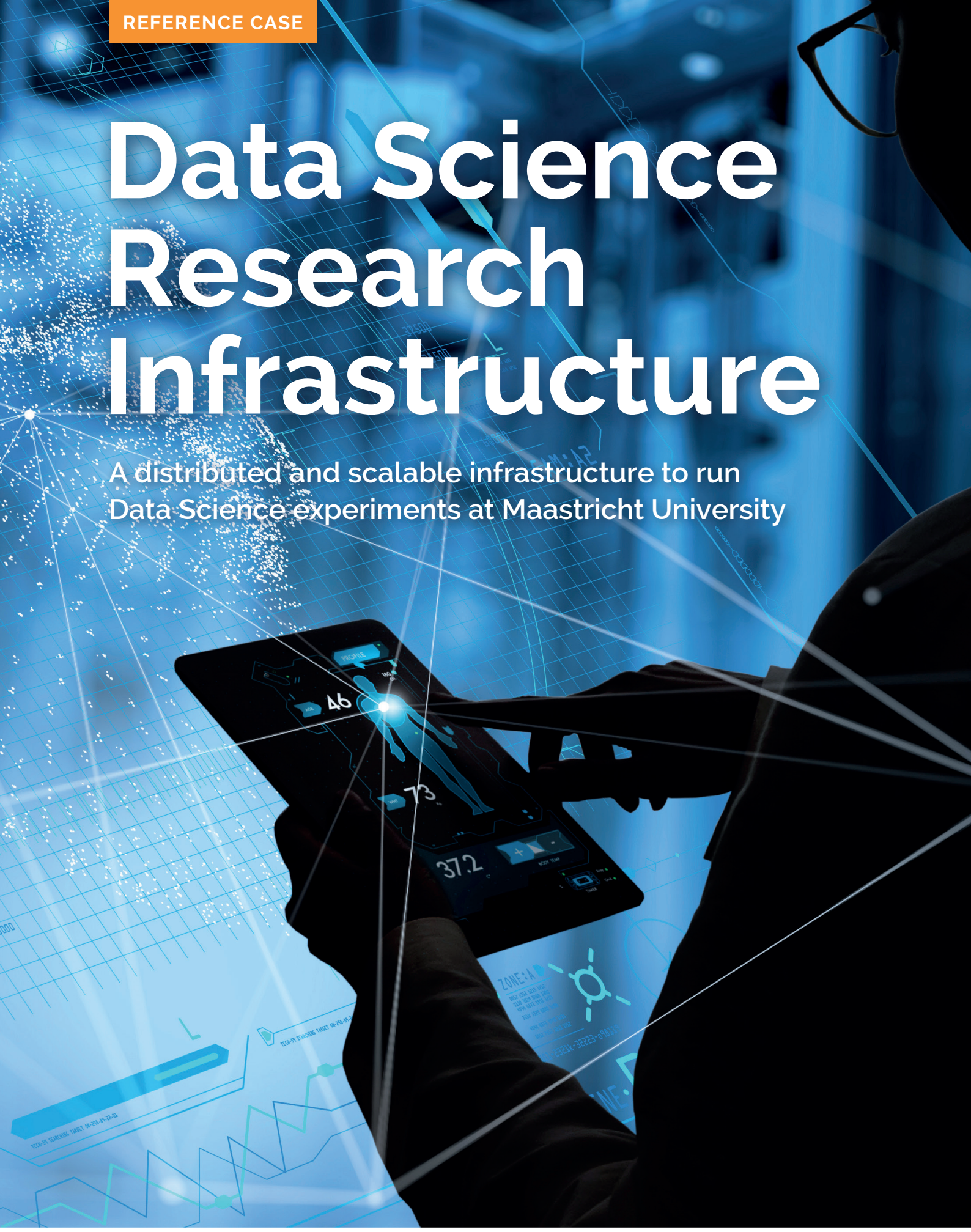
REFERENCE CASE

Data Science Research Infrastructure

A distributed and scalable infrastructure to run Data Science experiments at Maastricht University



Maastricht University FOURCO NVIDIA DELL Technologies

The work of data scientists can be computationally demanding and difficult to setup, necessitating a scalable infrastructure. If you happen to work at Maastricht University, you can use the Data Science Research Infrastructure (DSRI) for your research. Together with Dell Technologies, FourCo, NVIDIA and UM (Maastricht University) we created a unique research platform to facilitate the increasing use of Machine Learning and AI for research purposes and the need for on demand GPU processing power. An inspirational story.

The inspiration for the DSRI

Michel Dumontier is Distinguished Professor of Data Science and the Founder and Director of the Institute of Data Science (IDS) at Maastricht University.

The IDS develops and applies computational methods towards a wide range of societally relevant issues in medicine, the social sciences, law, and business. Researchers at the IDS work in teams with UM and external partners to develop their technology in the face of social, legal and ethical concerns, resulting in societally acceptable innovations. Crucially, the IDS also advocates for open science FAIR (Findable, Accessible, Interoperable, and Reusable) technologies. Over 30 people work at the IDS: 6 faculty members are joined by staff, post doctoral fellows, and PhD students.

As a PhD student, Michel helped deploy a 100 node cluster using fiberoptic networking and state of the art distributing computing techniques. That was over 20 years ago.
Since then he realised that this way of working is not only very useful for working with big data, but also that the work is far more manageable.

"It is also easy to use for people who are a little less tech-savvy, which lowers the barrier."

"One of the challenges that the IDS had, was to not only facilitate computing, but also put up services that would be available for other people to use, like database services. This is different from just a high-performance computing facility, like a 1.000 node cluster with scheduling. Hundreds of CPU's are really good for big problems, but most people do not have those big problems. They need something that's a bit better than their laptop and they need it for a short period of time. They also need services as part of their publication. That's a big gap. You can rent cloud service, but that will end up being very expensive once your service becomes very popular. It is important to leverage the infrastructure of the university to support the deployment, where you don't pay for data traffic.
At the same time we have a lot of people getting into data science. So we are training new data scientists and we want to make it real easy for them to get to work. They should not have to worry about configuring all of their packages, the dependencies and basic framework. This would simply take up too much of their time. We want to offer people something that makes it easier to deploy their own service instead of learning all the tools to do it themselves. This was the inspiration for the DSRI, the Data Science Research Infrastructure. The motto of the DSRI is: 'Deploy anything for free with powerful computer services'. It is a semi-managed middleware. People have access to a catalogue of well-defined services that are common to the group of data scientists. We also support them in making their project accessible on the infrastructure. We train them where they need to be trained, so they can use our services to deploy and get to work, without having to make virtual machines on their laptops and all those kinds of things."

Central infrastructure saves time

"The other benefit is that there now is a central infrastructure to do this, so we save a lot of time because data scientists do not have to build their own infrastructure for only their own research while also having to maintain it. This means fewer servers under desks or in closets, which also have to be managed and kept up-to-date and secure. That is administratively expensive and prone to failure. Instead, we now have a central solution that reduces the administrative burden of interacting with users.
It is scalable, so research groups can use it, but also students that need computing power beyond their own laptop."

From his first job interview, the topic of central data science infrastructure has been there for Michel Dumontier.
From local servers to high performance computing: there was a menu of options. The idea was to explore whether this was an opportunity not just for a little group, but for other groups at the university as well.

"We reached out to FourCo to get their advice on setting up the infrastructure and the software stack. They have a lot of experience in working with these kinds of technologies, so they were able to advise us on the procurement of the hardware and also liaise us with hardware vendors like NVIDIA, Dell Technologies and others for the procurement process. But FourCo was also physically on site when we were unpacking everything and trying to get it all set up."

"Even when the DSRI was already set up, FourCo was still available to us for consulting."

"We picked an open source software stack, the OKD OpenShift. The 3.11 OKD version was not easily configurable with our hardware. So it was important to have somebody who could guide our IT team to configure the cluster, to troubleshoot and to also interact with the vendor.
FourCo had similar clients, so they had experience to draw from. When you work with new hardware and open source software that is being developed, there are challenges.

Data Science Research Infrastructure

When those gaps become visible, it is good to have another group of people to discuss and find the right solution to the problem. And they have been great and very supportive in our effort to set this up. Even when the DSRI was already set up, FourCo was still available to us for consulting."

The DRSI project was pitched to the university in 2018, procurement was completed in early 2019 and the infrastructure was up and running in mid/late 2019. After that, there was a migration to OKD 4.6, which offers a lot more functionalities. In June 2020 the DSRI was running stable, so it was opened up to other users. In about a year and a half it grew from 10 to nearly 200 users and is expected to grow to several hundreds of users. These are researchers across all UM faculties with very different backgrounds, expertises and research questions.

"We invited them to get to know the DSRI and organised workshops to train them. We also provide project based support while they work on their projects. Our objective is to make sure everyone here at the university who does computing work is aware of the DSRI. We want to create a community and be more connected with other faculties."

The DSRI is a 16 node Dell EMC Cluster based on AMD EPYC CPU's with 64 cores 512GB of RAM 120 TB storage per node. This means a total capacity of 1024 cores and 1920 TB. The GPU node we use is an NVIDIA DGX1 8x Tesla V100 and 32 GB memory on each.

"By using the DSRI we can build better models because of the resources, computing power and speed."

Quality of research increases

These are the benefits that the DSRI offers Maastricht University:

- A stronger role for IT in providing solutions for researchers. They cater to the needs of researchers by making customized software packages and having discussions about code and security.
- Maastricht University has the vision to make its research data FAIR: Findable, Acceptable, Interoperable and Reusable. This objective is embedded in the DSRI, so the university is better positioned to achieve their open research requirements.
- The central infrastructure created a community of data scientists, which used to work spread across the different faculties, divisions and departments. Working together on it makes it better for everyone. It advances the research that all data scientists do. The objective is to enrich the academic culture around computing at the university. Celebrate successes together, use the DSRI for education of data scientists and also reach out to researchers outside Maastricht University in the future.

"We think the quality of research will increase because of the DSRI. Data scientists now have the right tools to package their data. There will be more open science, more transparency. It is a real opportunity. We like to onboard more people, even if they do not have the technical expertise, since we can train them. We will save them time, because the infrastructure they need is already there. It is all about efficiency and community. We hope more and more exciting applications will emerge. Time will tell."

Setting up the DSRI

The team at ICT Service at Maastricht University founded the DSRI together with experts from FourCo. But this empty tool of computing resources needed to be set-up.
Data science developer Vincent Emonet's role was mainly to find out how to deploy the right tools that make it easy to use these resources. He wanted to make the process easy for other researchers: "The learning curve can be a little bit steep at the beginning, especially for people who do not have a lot of experience in data science, and depending on what you want to do. But a lot of people told us they are really happy with it, because the DSRI makes it much easier for them. They are not limited by their own computers anymore when running big algorithms due to the DSRI. The reception is quite good."

"The aim of my research work is to represent this data optimally and to make it available to as many people as possible. With an emphasis on integration of data sets from different systems. We have a lot of computing to do, therefore part of my job is related to server administration, which enables researchers to focus on their research. I do support, and train users and students how to use the DSRI. The more I discuss things with other researchers, the more I realize that we have exactly the same problem, while coming from different backgrounds/domains."

"The DSRI brings a lot of freedom in terms of collaboration."

"For example in societal and history related research, such as ancient games, there has been some help from the things we learned by doing biomedical research. It is nice to have a common tool to share our knowledge. Biomedical researchers are slightly ahead of other domains in data science, but we are also pretty focused, so it is really nice to work with new people with fresh ideas. Interesting paths that we inititally did not think of, because we were already too deep into our own ways. Working on the DSRI is not set in stone, it is a journey we are all on together. Processing the data and extracting knowledge from the data is really hard to completely master. It depends on a lot of factors. The DSRI helps us to share this knowledge and tackle challenges we all face in our experiment."

Data Science Research Infrastructure

"If I would have run this on my laptop, which is pretty powerful, it would have taken a year to finish. With the DSRI I can get it done in a day."

"The DSRI brings a lot of freedom in terms of collaboration. It allows us to use more powerful resources, with more freedom. Before the DSRI was initiated, I was one of the rare person to have access to something more powerful than a laptop, but now most researchers and students at Maastricht University can access and leverage those resources. One of the advantages is that on the DSRI you use a lot of tools that data scientists are already using on their own laptop. You access them in your browser, write code and see the results directly. It is open source software, so when anybody wants to reproduce the experiment, they don't have to buy a licence to do it."

What do users think of the DSRI?

Suraj Pai, Masters in Artificial Intelligence Student at Maastricht University

"I'm in my first year of my PhD, I started in September 2021. Before that I used the DSRI when I was a Masters student. I worked at Maastro Clinic on the very popular field of artificial intelligence called deep learning applied to medical imaging. This is a radiotherapy clinic, which is a part of Maastricht University. We worked on different use cases applying deep learning to radiotherapy applications. One was about segmentation of tumor and organs at risk, so you can optimise radiotherapy planning helping save time for radiotherapists. Another one was trying to enhance the quality of a certain type of images called CBCT images, which are generally used for setting up patients during radiotherapy treatment. By optimising this image you can optimise the treatment process itself. By using the DSRI we can build better models because of the resources, computing power and speed."

"The support we get from the DSRI team and everyone that is involved has been great from day one. Anything we need or if there is a problem we encounter, within a day we would get a reply. Another advantage of the DSRI is that it is much easier to use than the alternate cluster system. You can run a development environment, which feels like you are coding on your laptop. The onboarding process is alsosuper simple. The state of the art GPU's are great as well, super fast and deliver very high performance. Since we work with 3D imaging, CT scans and MRI scans, we demand a lot more from the the GPU's, which the DSRI delivers."

Michiel Adriaens, assistant professor Maastricht Center for Systems Biology

"My research deals with the early detection and better understanding of diseases in which metabolism is somehow impacted, such as type 2 diabetes, but also in healthy people who do not live as healthily as the maybe could. In doing so, I use all kinds of techniques, including mechanistic models, that precisely simulate a small piece of biology. I combine that with data-driven methods, in which I look purely at the data that is available on a particular subject. This involves uncovering patterns in the data and making predictions based on them, thereby identifying patients or specific groups of people to look at more closely."

"For both approaches, I have a lot of use for the DSRI. Humans are very complex and as a result, there is a lot of noise and variation, which leads to requiring more measurements and therefore larger data sets. To process those, you need a lot of data storage and computational capacity. Before the DSRI, this was a lot more difficult. Back in 2015, we would buy our own specific computational servers, which was pretty common at the time; each department had its own equipment. The moment you buy it, it is already outdated, because there's never enough budget. For some tasks, such a server was certainly adequate, but it was not fast. You turned something on and a week later you looked to see if it was ready."

"Right from the start, the user-friendliness of the DSRI was huge when I compare it to our old server system, where we were limited to using a command line interface. Now you log in through your web browser and get a nice graphical interface, you select which software you need and you put the DSRI to work. It is also easy to use for people who are a little less tech-savvy, which in my opinion significantly lowers the barrier. In addition, there is now plenty of computing power for people like me who need it in their research. I am very satisfied with the DSRI."

"Besides the system and the technology, I think it is also important not to forget the people behind it. There is a whole team at the IDS and ICTS that is working on this. They have set up a very nice system around support: you know who can help you and these people are easily reachable through Slack and other channels. Additionally, it is really a community effort, with users from other faculties also helping out one another. I think that is a unique and fantastic setup. "

Data Science Research Infrastructure

Wolfgang Viechtbauer, associate professor of methodology and statistics in the Department of Psychiatry and Neuropsychology and the School for Mental Health and Neuroscience at Maastricht University

"I'm a statistician by training, applying, developing and examining methods for the analysis of data, for various research fields, like the social sciences and health sciences. I also do quite a bit of software development, which is an important part of my work. I write packages for R, that extend the capacity of R in various ways. The packages that I write for R are being used as part of the simulation studies on the DSRI."

"For data analysis, I run simulation studies when I want to know how good a particular method works. Is it precise or is there a bias in the method, so it overestimates or underestimates the generated quantity? I research this in situations where we know the truth and then examine how well the method works. We repeat this process thousands of times to get the properties of the method, which is really demanding. It can take easily weeks or months on your own computer, depending on how complex the simulation study is.
This is where computing infrastructure comes into play. More power means you can get the results more quickly. For me this is one of the main uses of the DSRI. I can run many simulations simultaneously. I was just running a simulation study which involved thousands of different conditions.
If I would have run this on my laptop, which is pretty powerful, it would have taken a year to finish. With the DSRI I can get it done in a day. That completely changes my ability to get my work done."

"I am super glad that the DSRI is now available at Maastricht University. It is a low barrier system which everybody can use. It shows that the university is willing to invest in high performance computing infrastructure. They recognize that it makes everyone's work a lot easier. That is important to me as well."

Want to learn more?

Visit the DSRI website: www.maastrichtuniversity.nl/dsri or send an email to dsri-support-l@maastrichtuniversity.nl (DSRI) or info-ids@maastrichtuniversity.nl (IDS).
Interested in modern IT infrastructure and data solutions? Please visit www.fourco.nl or mail us at info@fourco.nl

About Maastricht University Institute of Data Science
Founded in 2017, the Institute of Data Science (IDS) is committed to research in data science and artificial intelligence, collaborating across disciplines, institutions, and sectors.

About FourCo
FourCo helps your organization in making the right choices in IT Platform/Infrastructure and take ownership of the delivery of these choices. FourCo is located in Amersfoort, The Netherlands.

About NVIDIA
NVIDIA was founded in 1993, and is a leading manufacturer of high-end graphics processing units (GPUs). NVIDIA is headquartered in Santa Clara, California.

About Dell Technologies
Dell Technologies (NYSE:DELL) helps organizations and individuals build their digital future and transform how they work, live and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the data era.

Maastricht University

FOURCO

NVIDIA

DELL Technologies