

Sydney Informatics Hub Research data analysis and training

Sydney Informatics Hub enables excellence in computational and data-driven research by providing support, training and expertise in statistics, data science, AI, software engineering, simulation, visualisation, bioinformatics and research computing to University researchers and external collaborators.



Find out more

sydney.edu.au/sydney-informatics-hub

New gateways to agricultural data

Challenge

Australian agriculture is valued at \$71 billion dollars and relies on data to drive optimisation. In 2021 with funding from the Australian Research Data Commons (ARDC), the Agricultural Research Federation (AgReFed) platform was launched to share agricultural data and benefit the sector. The project team was confronted with the task of managing and extracting insights from vast, diverse sets of spatiotemporal data such as climate, satellite and soil observations. Existing manual processes were time-consuming and difficult to reproduce, hampering potential for machine learning and other advanced analytical approaches. A rethink of the architecture was needed.

Research

Led by Dr Sebastian Haan, Sydney Informatics Hub data scientists and high-performance computing experts developed two powerful reusable tools, Geodata-Harvester and the AgReFed Machine Learning Model, to improve accessibility. The Geodata-Harvester tool provides open-source software and machine learning to easily extract and align data for specific regions and time periods, while its counterpart uses Al to map environmental properties and their uncertainties from spatial and temporal data.



Results

The tools are now freely available online, enabling access to more intuitively aggregated geospatial, climate and soil data for use in research and decision-making. The software has been published as open-source packages, allowing other developers to build upon the models, and the outcomes have led to further funded research engagement with AgReFed.

Funding sources include:

- Australian Research Data Commons
- Grains Research and Development Corporation



Above: Concept map for the AgReFed Geodata-Harvester AI tool, developed by Sydney Informatics Hub.

Improving children's oral health

Challenge

Poor oral health can impact a child's development, and in adulthood is associated with chronic disease. Dr Christina Adler (School of Dentistry, Faculty of Medicine and Health) and her team designed a ground-breaking study to examine the links between oral microbiome development and oral health in childhood. To complete the study, the team needed a process for analysing longitudinal data from over 200 children to identify antimicrobial resistance in each child's microbiota and its influence on tooth decay.

Research

Sydney Informatics Hub drew on expertise across three teams to offer a data analysis solution. First, Dr Cali Willet from the bioinformatics team built a scalable bioinformatics workflow to create data on the relative abundance of different microbiota. Statistical consultants including Dr Kathrin Schemann then enabled the research group to develop their analysis plan and



implement advanced statistical modelling, leading to better and more precise model parameter estimates. Finally, analysis by Dr Henry Lydecker and the data science team generated statistical insights into the diversity and temporal evolution of microbiota and antimicrobial resistance, as well as heritability, composition and impacts on tooth decay.

Results

Sydney Informatics Hub's data analysis support enabled the completion of a successful \$4.3m US National Institutes of Health grant application, high-impact publications including a 2023 paper in Nature Communications, and promising findings for the development of probiotics to improve children's oral health. The study has also advanced understanding of links between the oral microbiome and antimicrobial resistance, which is a major threat to human and animal health globally.

Funding sources include:

- Channel 7 Children's Research Foundation
- Financial Markets Foundation for Children
- Intersect
- National Health and Medical Research Council
- National Institute of Dental and Craniofacial Research
- Multiple Births Association
- Twin Research Australia

Left: Dr Christina Adler, School of Dentistry, Faculty of Medicine and Health.





Training Upskilling researchers

On top of its data analysis services, Sydney Informatics Hub nurtures research excellence through training in advanced analytical techniques. Its rich program shares the expertise of its highly skilled data specialists in short, modular workshops with downloadable course content.

In 2022, Sydney Informatics Hub ran 64 workshops, upskilling 1100 researchers across the University with over 95% satisfaction from the attendees.

695

researchers upskilled

sydney.edu.au/ informatics-hub/training







95+%

workshops

satisfaction

Accelerating wildlife conservation research

Challenge

Australia has more than 500 threatened animal species, and non-invasive wildlife monitoring with motion-activated cameras is crucial to their conservation. However, a single study can generate millions of images, and identifying the animal in each picture is time-consuming and costly. Dr Aaron Greenville (School of Life and Environmental Sciences, Faculty of Science) and NSW National Parks were keen to develop a fast, efficient way of identifying animals in images to maximise the benefit of camera data for wildlife conservation and management.

Research

Data scientists including Dr Henry Lydecker at Sydney Informatics Hub set out to create Marsupial.Al, a visual Al prediction tool to support automatic identification of 72 native and invasive species. They trained a state-of-the-art visual Al model using over two million labelled images of Australian wildlife from the dataset of WildCount, NSW National Parks' long-term animal monitoring program.

Results

Marsupial.Al can label tens of thousands of images per hour, compared with an estimated 450 images per hour for experts and around half that for citizen scientists. Precision is high at 94% for the most common half of species and 88% overall. With results 1,000 times faster and 5,000 times cheaper than manual approaches, it's a game-changer for Australian conservation research. Says Dr Greenville, "The time gained can be used for on-ground species conservation management" – a win for all.







