Dr Thomas Etherington on The Future of Digital Research

Geoinformatician / Ecologist at Manaaki Whenua - Landcare Research

Drivers of research

Please describe the global and local context for your communities' research over the next 5-7 years. Are there any linkages you see might emerge over that time, which might have a significant influence on our progress as a nation? What are the main drivers that shape the types of research being done? (industry drivers, government support, new product development, theoretical)

Having just returned to NZ from the UK, this is a tricky question. At Manaaki-Whenua our funding primarily comes from the government, so government priorities will influence what we work on. There is a focus on producing more applied research for the public good, so our work will largely be driven by events (that are somewhat unpredictable) in society and the environment. An example would be trying to understand the invasive potential of non-native conifers under climate change.

Research methods

In a broad sense, can you describe the research priorities you and your colleagues will be working on in 5-7 years? What do you see will be the fundamental differences in the types, complexities, and scales of research problems your community will be working on in 5-7 years?

In a broad sense, I'm not sure my research priorities will change much going forward. Fundamentally I will remain concerned about questions of where things are and how they are connected. Sudden and unknown events, such as disease outbreaks like Myrtle rust, will suddenly refocus exactly what we want to know the distribution of and its movement around a landscape. What will change is the type and scale of data that we will use. I can see a shift from smaller scale bespoke designed studies, towards making use of big data. While I think well designed bespoke longitudinal studies/surveys are very important for properly monitoring long-term change, they are limited in scale and may not necessarily occur in a location that ultimately becomes of interest.

Big data has an advantage here in that large amounts of unstructured data can be collected quite quickly. Citizen scientists can use smart phones to rapidly collect observations about the natural world, and can deposit them into databases such as iNaturalist where the data can then be freely accessed by scientists. However, turning this unstructured data into information and knowledge is a new challenge, as without a robust study design it becomes much harder to assume things about our data, and as such, conventional methods for data analyses may be inappropriate and hence potentially misleading. So I think that the research community will expand effort in examining just how much we can rely on big data – and this is certainly something I am hoping to work on myself!

Collaboration

Can you describe how much collaboration would be needed in your community both internationally and in NZ, in order to achieve your community goals in 5-7 years?
Also, what sorts of cross-discipline collaboration would be needed in 5-7 years to address the problems you want to solve?

I suspect that there is a growing recognition that much of the ecological big data we are using is as much the result of a social process as an ecological process, and as such the data cannot be viewed without engaging with those social aspects. Therefore, I can foresee myself working more with social scientists to try and connect the social processes that they study and ecological processes that I study to get a better overall picture/understanding. Beyond that, I can just see a greater need to engage with computational scientists and research software engineers. While I consider myself a very competent computational scientist, I am well aware of my limitations and that there are occasions when I would greatly benefit from computational experts. We have hired our first research software engineer, and this has been my first opportunity to work alongside such an expert. The potential for more efficient and better research seems quite obvious, but there are challenges of language/approach/terminology that have made the initial process somewhat of a challenge. I'm not sure how you resolve those sorts of cross-domain misunderstandings and confusion, other than to recognise that they will be there, and that all parties need to feel happy to ask "stupid" questions, and that everyone tries to be clear about what they are saying.

Skills and Capabilities

We are interested to understand whether researchers in your communities will be well positioned to solve the problems they want to solve. Does the community have a clear direction for developing its capabilities, and either way which capabilities will be essential to meet their medium-term to long-term needs?

This is a trickier one for me to grasp, having just returned to NZ from the UK, so what I will say here relates more to the UK situation. In general I think that Universities are getting better at incorporating computational skills in their degree programs. For example, I taught a programming course as part of a NZ Masters in GIS program, and I think that was well received by most of the students. I was also aware that in the UK there was a real growth in opportunities to study very quantitative/data science degrees. I'm not sure that this is the case yet here in NZ, perhaps because there isn't the critical mass of students wanting to specialise in geoinformatics to warrant such a program.

I think what is perhaps more of a problem in the next 5-7 years is upskilling those people already in research positions. There is some research that shows that in the recent past computational skills were not a strong part of degree programmes, and this means that many institutions will have staff who lack the necessary skills. So I think a major challenge may be upskilling those people already in the workforce who did not receive sufficient training to be able to continually learn and keep up to date with things.

Adoption of Research Computing

We're looking to understand the relevance and importance of advanced research computing to your community in the future. In 5-7 years how prevalent do you see computational skills as being across your communities? On that same time frame, can you foresee any other advanced digital research skills and methods which will be fundamental to achieving quality and impact in your research areas?

I would certainly like to see advanced research computing used more. Whether that happens or not will probably be a result of training and access limitations. HPC is fundamental due to data size. Meanwhile, the models being used will remain fairly simple. From conversations with other scientists, I think many still perceive HPC as 'too hard' and may prefer to ask simpler questions that can still be achieved on a desktop workstation. I'm not sure how such a barrier can be overcome, other than continuing to demonstrate what can be done, and keeping the barriers to doing this as low as possible. Those barriers need not only to relate to working on the cluster itself, but also to the dull logistical things such as registering for user accounts. I think the logistics are something that is often overlooked, as regular users will have perhaps only gone through that process once years ago. But for those people trying to decide if working on HPC is worth the effort, I would suspect that if the process of signing up is hard or convoluted (and to be honest I found that to be a slightly irksome process myself) then I think there is a real possibility of losing people before they have even got access.

I think the "*to achieving quality and impact*" bit of the question is perhaps more interesting. A little knowledge/skill can be a dangerous thing, and I would wonder just how well people are doing computational experiments. To achieve quality and impact the research should ideally be highly reproducible, but my experience of looking at work in my role as journal reviewer and editor would suggest that the quality can be highly suspect. I wonder if this is something like going beyond simply "hacking" together code that gives you an answer, to then going the extra mile to properly document things such that computational work stands a good chance of (a) being correct and therefore providing quality, and (b) being reused and therefore providing impact. I guess this links back somewhat to the training questions, but I'm not sure many people see the value of reproducibility once they have "their answer".

What can NeSI do?
What can NeSI do across the next 5-7 years to lower barriers, and increase productivity and excellence?

Access, training, and support! If I can't access the compute resources I'm obviously not going to use them. Personally I think I like a model that makes the access as open as possible, but with priority towards staff from those organisations that fund NeSI or that have providing direct funding for their compute time. I'm pretty sure this is how things work, but I would like to think that anyone should be able to work on the cluster if there is spare resource.

Supporting training would also be very helpful. Clearly there needs to be training on how to use NeSI systems, but providing more general community support through efforts like Software and Data Carpentry workshops I think is very helpful. The focus from NeSI should probably be more towards the Software Carpentry workshops, as I would imagine domain experts would be better suited to deliver Data Carpentry style workshops. Hopefully if NeSI staff can put in enough initial effort, a community of trainers will grow and things will become a little more self-sustaining.

Ongoing support is also key. NeSI providing experts who can facilitate people using HPC and optimise code is very important - someone to help with shell script to distribute the jobs evenly, for example.