Machine learning infrastructure deployed at scale: understanding requirements, demand, impact and international best practice

Wojtek Goscinski, Komathy Padmanabhan

Monash University, Research Computing Centre (RCC) at The University of Queensland, Alfred Health, Victorian Institute for Forensic Medicine, University of Auckland, NVIDIA

Australian Research Data Commons

This research/project is supported by the Australian Research Data Commons (ARDC) The ARDC is enabled by NCRIS

Our Study

Our goals for the study were:

1. To form a clear understanding of the relationship between research requirements, computing capability, capacity, and research impact.
2. To understand individual researcher requirements and consolidate these across a large cohort of groups, so that we can make evidence-based recommendations on how to underpin research adopting ML in the most efficient and effective manner, at scale.
3. To understand international best practices, and how it should inform Australian investment.

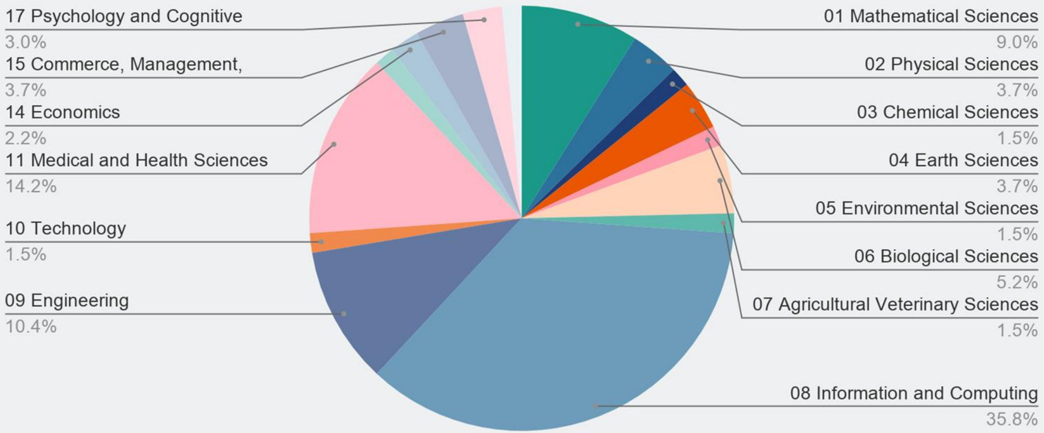
We identified **128 research groups** that apply or develop ML techniques, across Monash University, University of Queensland and University of Auckland.

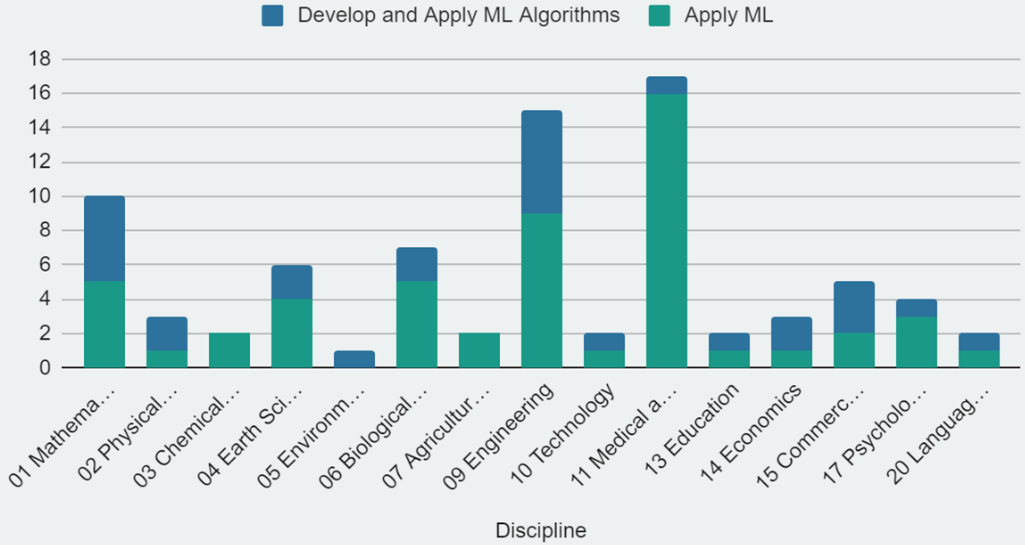
Of the identified research groups,
68 completed the survey (53% response rate).

The community is growing quickly as acknowledged by all three sites: MASSIVE at Monash, University of Queensland and University of Auckland.

Meetings with a number of international sites: Centre for Clinical Data Science, Partners Health, Boston, University of Michigan Advanced Research Computing & Michigan Institute of Data science, Jetstream at Indiana University, Cambridge University Research Computing, CERN

There is a strong and growing appetite across research groups for access to ML capacity, services, libraries, expertise and training.


17 Psychology and Cognitive
3.0%
15 Commerce, Management,
3.7%
14 Economics
2.2%
11 Medical and Health Sciences
14.2%
10 Technology
1.5%
09 Engineering
10.4%
01 Mathematical Sciences
9.0%
02 Physical Sciences
3.7%
03 Chemical Sciences
1.5%
04 Earth Sciences
3.7%
05 Environmental Sciences
1.5%
06 Biological Sciences
5.2%
07 Agricultural Veterinary Sciences
1.5%
08 Information and Computing
35.8%


Develop and Apply ML Algorithms
Apply ML
18
16
14
12
10
8
6
4
2
0
01 Mathema...
02 Physical...
03 Chemical...
04 Earth Sci...
05 Environm...
06 Biological...
07 Agricultur...
09 Engineering
10 Technology
11 Medical a...
13 Education
14 Economics
15 Commerc...
17 Psycholo...
20 Languag...
Discipline

Accessing Compute Capacity

Current computing capacity is highly inadequate with requirements growing quickly.

- **68% of the researchers expect that their compute needs to grow by 100%-200% in the next year.** 15% expect that their computing needs will grow by over 200%.
- **54% of the researchers indicate that compute capacity is their major challenge.**

Extrapolation across entire survey cohort of 128 research groups identified across participating Universities

Total GPU hours by these respondents	3,008,802	GPU hours
Size of GPU cluster required to accomodate	382	GPUs

Extrapolation across Australian Go8 (Assumes the three participating Universities are a representative sample of the Australian Go8 - 3 of the 8)

Total projects	341	projects
Total GPU hours by these respondents	8,023,471	GPU hours
Size of GPU cluster required to accomodate	1,018	GPUs

Infrastructure Recommendations

Develop investment and access models to allow national access to GPU computing.

Increase knowledge to improve efficiency of usage.

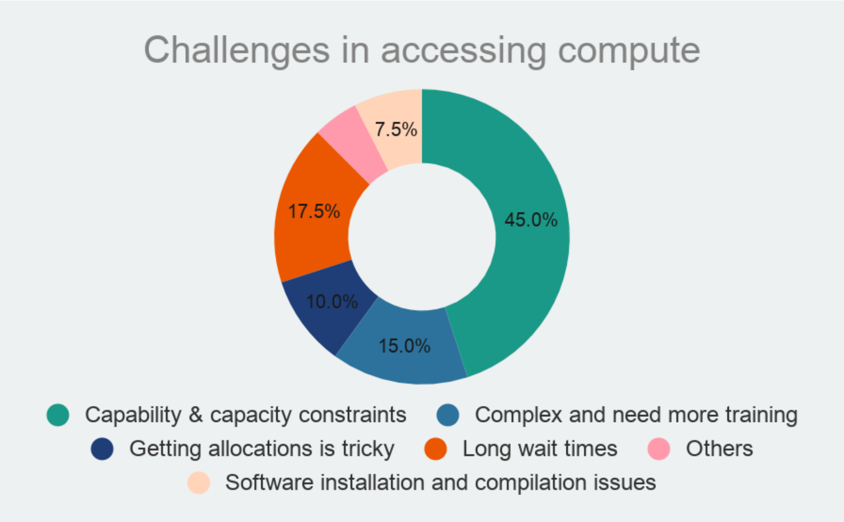
Provide knowledge, training and test environments to allow researchers to test ideas, demonstrate success and potentially fail quickly.

Increase system administrator knowledge on how to deploy and configure libraries efficiently and effectively to best leverage the limited hardware available.

Infrastructure providers should collaborate nationally to increase the overall pool of available resources. HPC centres offering significant GPU capacity are important to this effort as they represent the majority of available GPU capacity for research in Australia and are the primary support contact.

Computing Environments

63% of researchers use HPC facilities to access GPU capability. 19% use commercial clouds.



Commercial Cloud

19% of respondents indicated that they have now or in the past used commercial cloud for ML.
AWS and Azure are most commonly used.
Key advantages of commercial cloud highlighted:

- Quick access
- Cloud credits for research
- Availability of pre-trained models

Key challenges of commercial cloud highlighted:

- Very expensive to scale
- Memory not sufficient
- Scheduling is manual
- Complete setup process & error messages
- Short term free credits is not worth the effort of moving data and setting up models, since it is time consuming
- Not suitable for sensitive data due to ethics requirements & efforts in redacting the data
- Uncertainty about IP

Techniques and Environments

Neural networks & Deep Learning (79%) and Linear regression (32%) are the most commonly used ML techniques.

Tensorflow, Keras, Pytorch, scikit-learn and Caffe are the most widely used ML tools/libraries/frameworks.

Pandas for Python, Matlab and Moa are used widely for data manipulation, mining and analysis.

Currently researchers must move between environments to develop, train and test. This includes their personal desktop PC, HPC systems, clouds and storage, and this introduces inefficiencies.

Infrastructure Recommendations

The focus on key libraries such as Tensorflow, Pytorch and Caffe is an opportunity to develop and spread detailed knowledge across a wide variety of research challenges. It means **a large number of researchers can be assisted by providing expertise across a small number of tools/libraries/frameworks**.

Create **integrated infrastructure** that allow researchers to access interactive development environments, interactive compute for testing training, heavier compute for dedicated training, data manipulation tools, and access to data, in an efficient and integrated manner.

Data

Availability & accessibility of quality big data is a challenge for more than 50% of the researchers surveyed. Limited availability of annotated/labelled datasets means researchers spend a lot of time and effort pre-processing data.

A key highlight is the sensitivity of the data & privacy issues associated with it. **53% of the researchers indicated that their data is sensitive and needs to be secured and managed appropriately.**

A large proportion (73%) of the researchers use public reference datasets for training, testing and benchmarking their models.

Infrastructure Recommendations

Provide access to existing curated and well annotated data repositories for ML adopters and practitioners. Provide access to this data on ML hardware.

Integrate ML capability with the appropriate controls to host sensitive data to underpin the 50%+ of researchers who require strong data security and governance.

Curate and make available reference data sets commonly used by ML practitioners as a standard offering across GPU computing facilities.

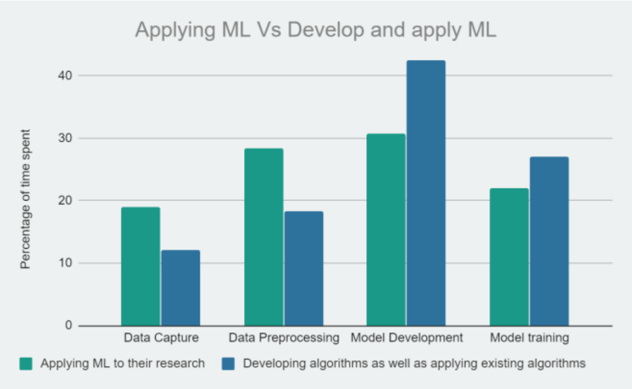
Knowledge and Training

The lack of researchers with machine learning skills and domain expertise is seen as one of the key challenges in applying ML to research.

Training needs are found to be different for ML researchers who develop algorithms and ML researchers who apply ML to their research.

Understandably, training needs are also found to be different for researchers with different expertise levels.

90% of ML researchers are multidisciplinary



Infrastructure Recommendations

Provide targeted training programs & hands on workshops to ML developer cohorts and applied ML researcher cohorts.

Provide training on research specific tools and IDEs like NVIDIA Clara for Healthcare researchers, PyTorch for computer vision and NLP researchers.

Promote ML as a major tool for multidisciplinary research to grow profile

Build specific national communities of practise in the identified areas by:

- Developing and promoting forums and exchange of ideas,
- Targeted training,
- Regular remote hybrid training.

Acknowledgements

Monash eResearch Centre, Monash University
Research Computing Centre (RCC) at The University of Queensland
Alfred Health
Victorian Institute for Forensic Medicine
University of Auckland
NVIDIA
Centre for Clinical Data science, Partners Health, Boston
University of Michigan Advanced Research Computing & Michigan Institute of Data science
Jetstream at Indiana University
Cambridge University Research Computing
European organisation for Nuclear Research, CERN

68 research group respondents across Australia and New Zealand

Australian Research Data Commons

This research/project is supported by the Australian Research Data Commons (ARDC) The ARDC is enabled by NCRIS