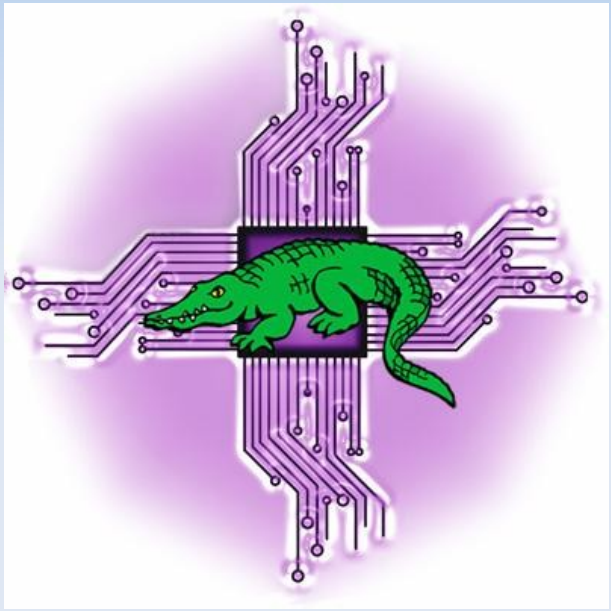
HPC Bioinformatics Education: The University of Melbourne Experience



eResearch New Zealand, February, 2025

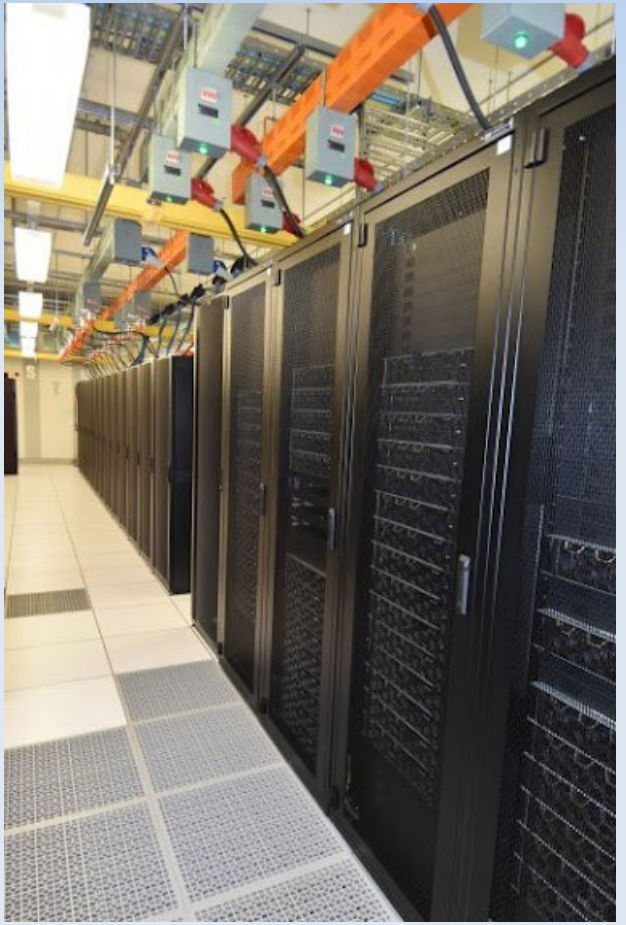
Lev Lafayette and Daniel Tosello

Research Computing Services, University of Melbourne

HPC Education: A Review

- **Many past presentations and papers about HPC education – and many presented in Aotearoa New Zealand.**

- **"Issues and Solutions in Teaching Researchers The Value and Use of High Performance Computing" (New Zealand eResearch Symposium 2011), "Teaching Scientists High Performance Computing" (Otago University Systems Research Group, 2013). "Critical Issues in the Teaching of High Performance Computing to Postgraduate Scientists" (ICCS, 2014). "Training and Education in High Performance Computing for eResearchers" (eResearch Australasia, 2014). "Skill Improvements versus Interface Designs for eResearchers" (eResearchNZ, 2015). "Software Tools Compared To User Education in High Performance Computing" (THETA, 2015). "Teaching High Throughput Computing: An International Comparison of Andragogical Techniques" (eResearchAustralasia, 2017). "HPC Case Study for Adult Learning Principles" (Australian National Data Service, 2018). "International HPC Certification Program" (ISC, 2018). "Towards an HPC Certification Program" (SC, 2018). "Towards an HPC Certification Program" (Journal of Computer Science Education, 2019). "The International HPC Certification Forum and AU-NZ" (ARDC, 2019). "One Year HPC Certification Forum in Retrospective" (Journal of Computer Science Education, 2020). "Training and Curriculum Development for International HPC Certification" (HPC Cert Forum, 2020). "Contributing To the International HPC Certification Forum" (eResearchAustralasia, 2020). "The International HPC Certification Forum : A Call for NZ Contribution" (eResearchNZ, 2022). "HPC Training Generates HPC Results" SCAsia, 2024), "HPC Certification Forum & Skill Tree: An Update" (SCAsia, 2024), "Tailoring HPC for Medical Research: Lessons from the Parkville Precinct" (eResearchAU, 2024)**



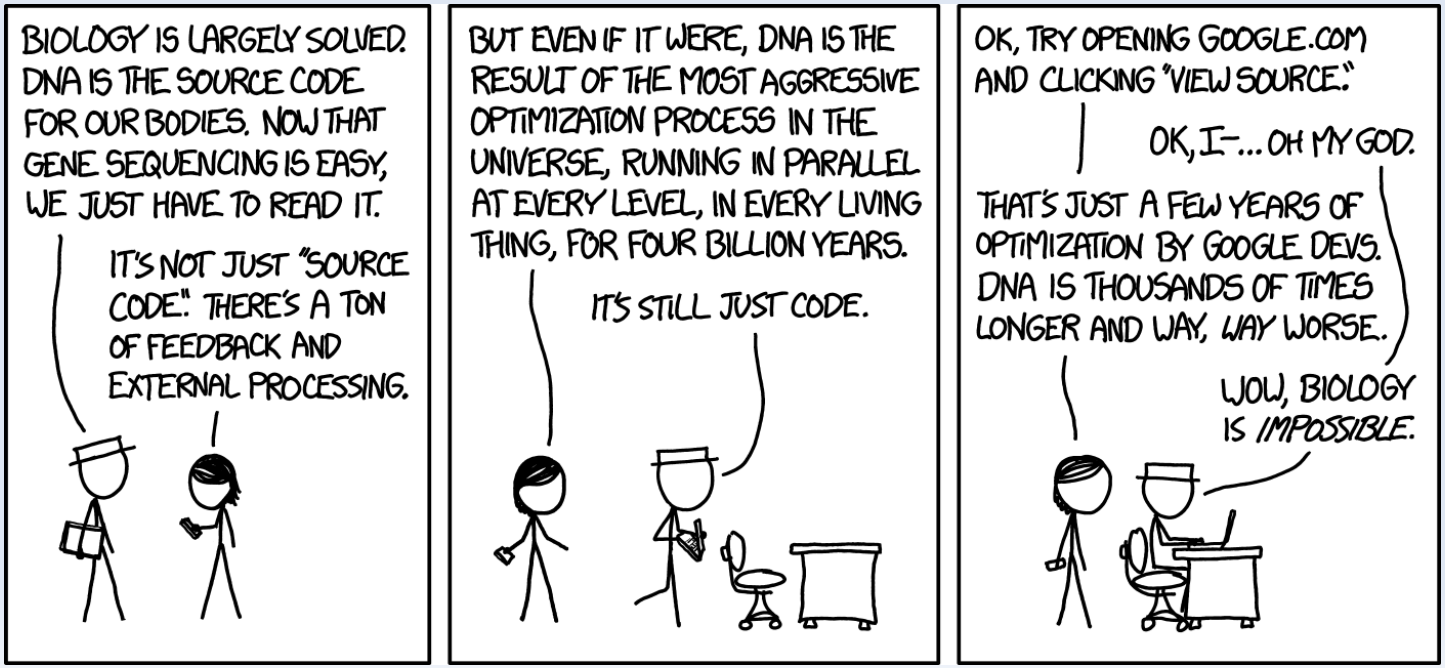
Bioinformatics at UniMelb

- **This presentation is largely descriptive; i.e., what we have and how we do it. However part of it identifies specific needs for bioinformaticians and how that is built in to our training programme for researchers. That might be useful for others! :)**
- **University of Melbourne and the Parkville precinct has a strong Medical Research Institute (MRI) presence. Figures below as of October 2024. MCRI and Peter Mac are undergoing refresh; figures are not exhaustive.**

	WEHI	PeterMac*	MCRI*	UniMelb
Nodes	98	26	17	181
CPUs	3,496	1,060	768	14,624
RAM (TB)	44.7	~9	~15	149
GPUs (NVIDIA)	20 P100 28 A30 6 A100 (40 GB) 4 A10	3 V100 6 T4	4 T4	124 A100 (80 GB) 40 H100 SXM5
Filesystem	VAST	GPFS	GPFS	GPFS

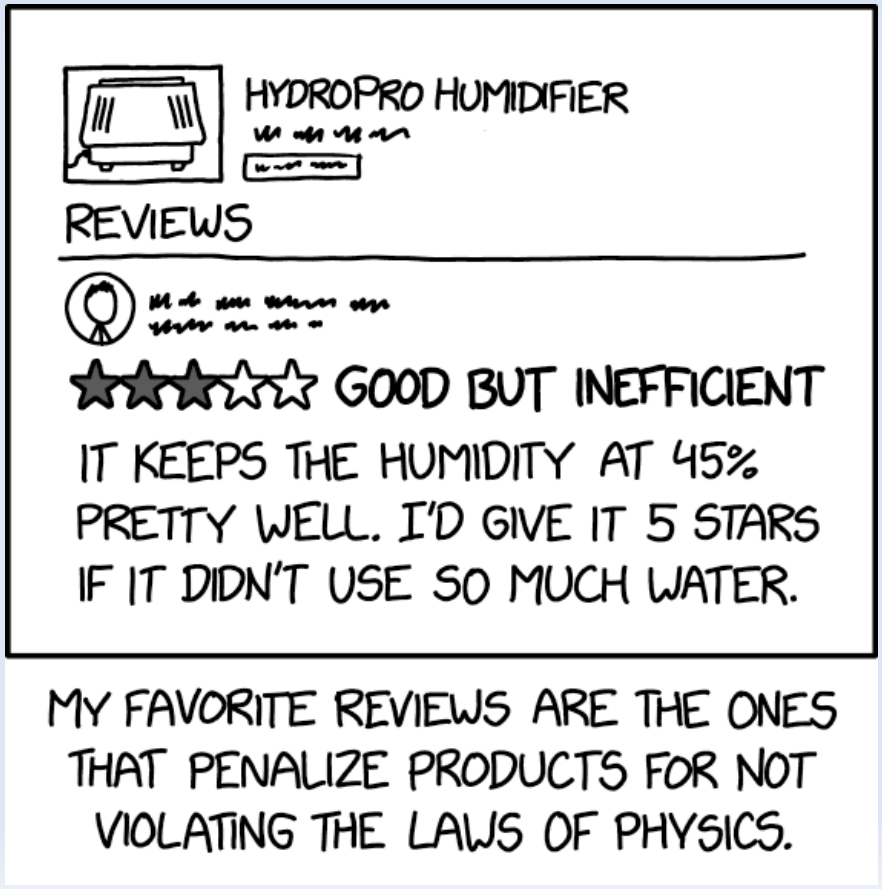
Bioinformatics: A Definition and UoM

- **Simple definition: The use of software for understanding large and complex biological data. Can include: sequence analysis (sequencing, assembly, annotation, genomics), proteomics, phenomics, transcriptomics, molecular modelling, cryogenic electron microscopy.**
- **UniMelb and precinct covers these fields extensively; Melbourne Bioinformatics (formally VLSCI), Doherty Institute, Walter and Eliza Hall Institute of Medical Research, Peter MacCallum Cancer Centre, Murdoch Children's Research Institute, etc. UniMelb has Schools of Biosciences, Biomedical Sciences, Health Sciences, Population and Global Health, Melbourne Medical School, and Chemical and Biomedical Engineering.**



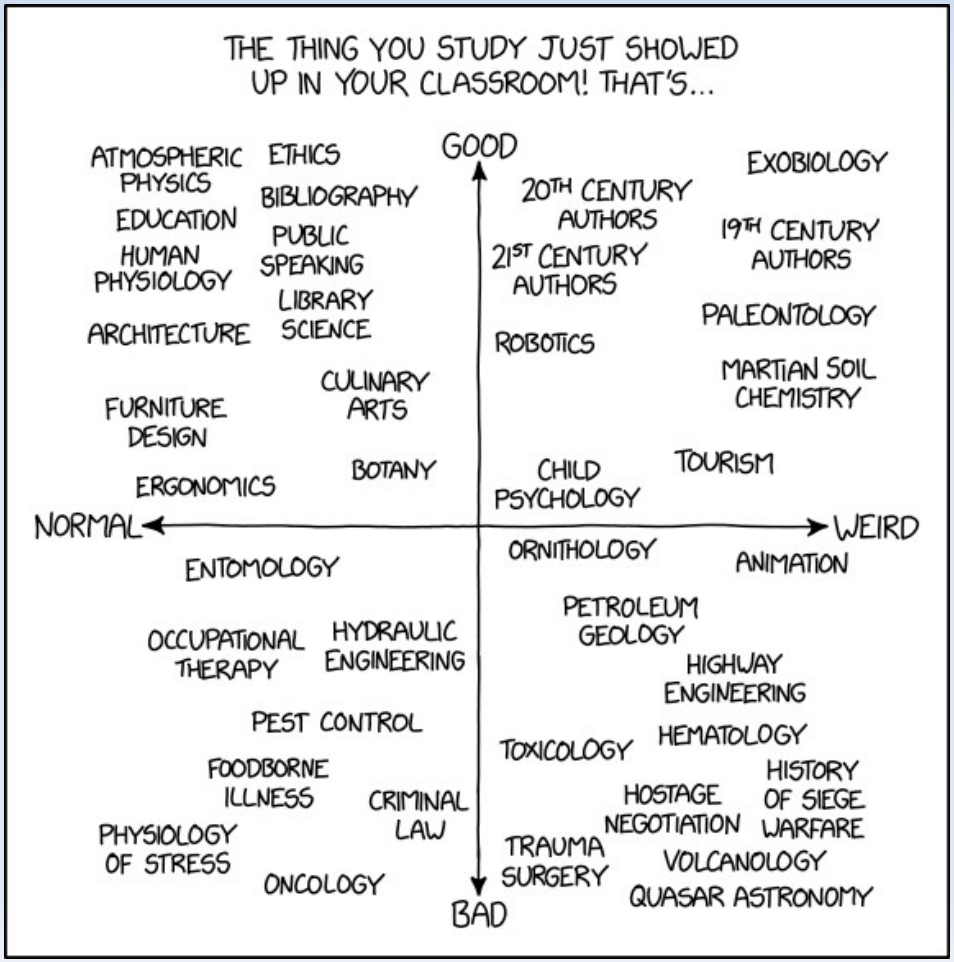
HPC Without Training

- **Many bioinformatics researchers are familiar with web-based systems (e.g., Galaxy https://usegalaxy.org/) and are unfamiliar with command-line tools, environment modules, HPC job submission scripts, etc.**
- **Tools like OpenOn Demand provide a lower barrier to immediate entry and can integrate familiar programming environments such as Jupyter and Rstudio. Nextflow is also commonly used for workflows and can be integrated into job submission scripts.**
- **However, these tools will *always* be suboptimal in terms of performance. A lower barrier to entry using applications cannot (i.e., it's physically impossible) perform better than a skilled user.**
- **Hugh Shanahan, Professor of Open Science, when asked the question:**
  "Shouldn't we build more modern tools than using 30 plus-year old commands?"
  "Oh get serious and grow up"
  -- ARDC Conference eResearch Skilled Workforce Summit, July 2019



Advanced Adult Education

- **Principles of adult education and advanced eduction apply for Bioinformatics researchers.**
- **Adult education; (1) autonomy of direction in learning (2) importance of personal experience as a resource (3) the emphasis on intrinsic rather than extrinsic motivations. Supplemented with "lifelong learning".**
- **Content needs to organised in terms of objectives, timed, and revised! Content needs to provided in as modular 'structural knowledge', with narrative, analogies. Grounding to a concept; facts and reasons provides understanding which allows further elaboration by the learner.**
- **Delivery should make use of discipline-based learning styles. For computer use, connectivism (e.g., paired programming) and direct usage ("yield to the hands-on imperative"). Needs to be followed up with feedback, and proximal learning with a follow-up mentoring and outreach program.**

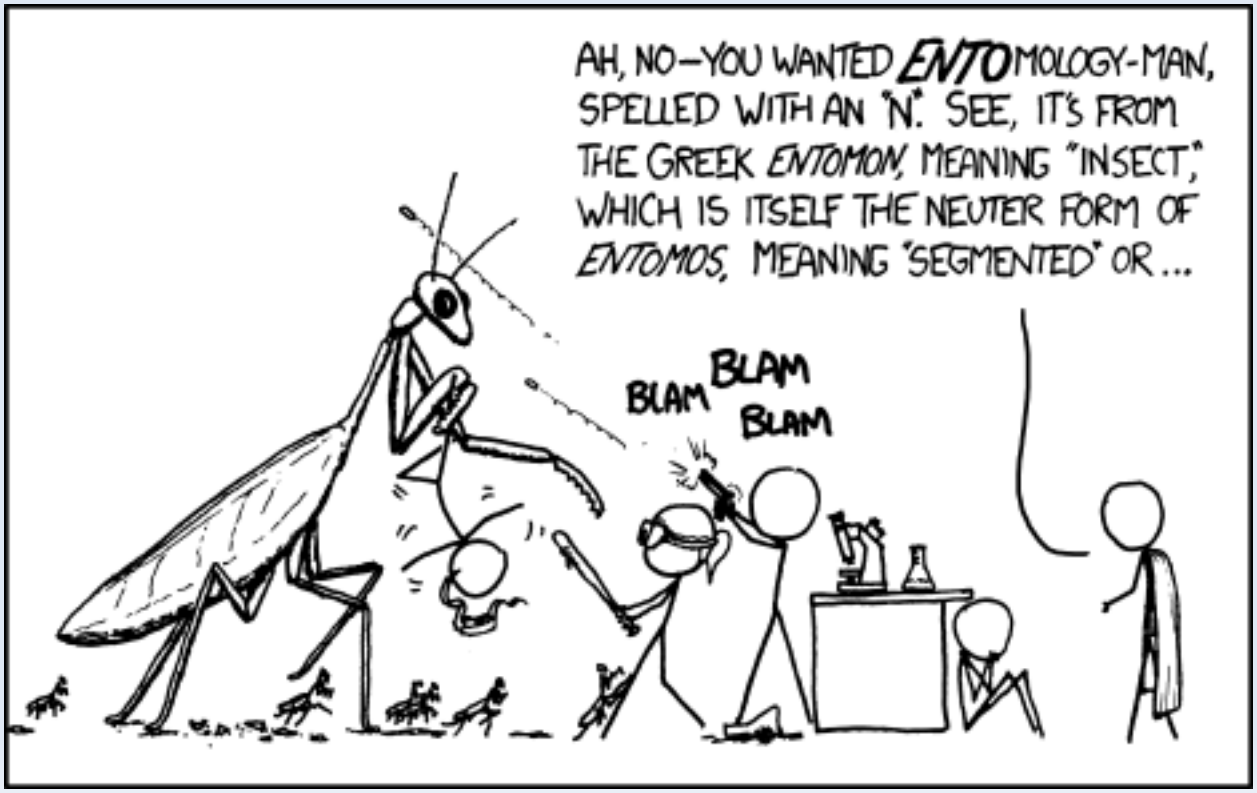


UniMelb HPC Training

- **The University of Melbourne runs regular day-length onboarding workshops. The two fundamental workshops are "Introduction to Linux and High Performance Computing" and "Advanced Linux and Shell Scripting for HPC". Like “software carpentry” but for HPC.**
- **With these recommended prerequisites there are several other workshops available: "Parallel and High Performance Python", "Regular Expressions on HPC", "Parallel Programming", "GPU Programming", "From Spartan to Gadi", "Mathematical Applications and Programming", "Data and Databases on HPC".**
- **Specialist versions of courses also exist for bioinformatics, mechanical engineering, and neuroscience.**
- **Workshops have been conducted in-person and via Zoom. Access is provided to text information, videos, and extensive job submission examples. Presentations are broken up into short modules which include formative spot questions and with extensive opportunities for researchers to elaborate on their own issues.**

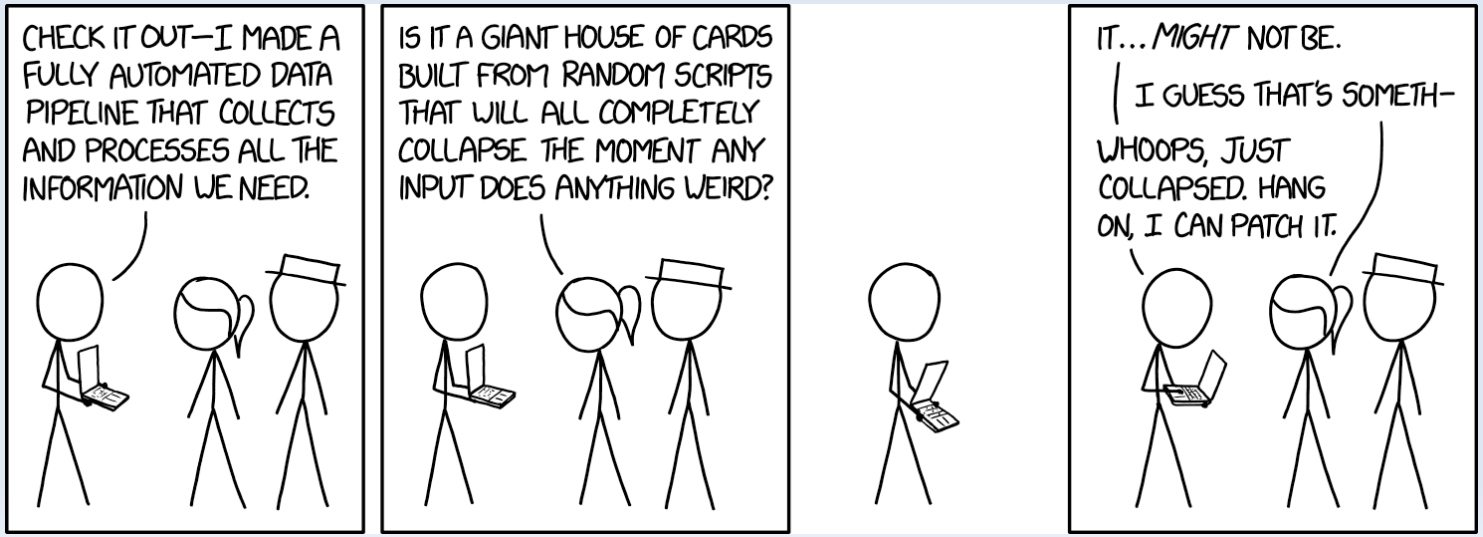
Spartan Champions Programme

- **The Spartan Champions programme exists so that experienced users in a project can provide onboarding and assist other members of their project team. In the process they gain access to a priority queue.**
- **The programme also includes monthly meetings with technical presentations and assistance.**
- **Thus there are four layers to HPC education at the University of Melbourne (a) no training, just access to documentation (b) HPC essentials, with tailored courses for particular disciplines, including bioinformatics (c) HPC specialist courses and (d) Spartan Champions programme.**
- **Approximately 50% from bioinformatics projects (e.g., Microbial eukaryotes, Direct RNA-seq on human-bacterial transcriptomes, species distribution modelling, Myalgic Encephalomyelitis)**



Bioinformatician Needs

- **Most of the HPC needs of a bioinformaticians are the same as researchers from other disciplines. They need exposure and familiarity with the architecture of the system, the operating system at a command-line level, environment modules, they need to understand partitions and queues, workload management, scheduler directives. More advanced users might also want to learn a wider selection of commands, regular expressions, shell scripting.**
- **Bioinformaticians have more particular needs as well, including: instrument to datastore, preprocessing; data management and data transfer; workflow management; use of GPUs; visualisation.**
- **UoM workshop includes a full workflow job submission from Data Carpentry wrapped in Slurm scripts for genomics along with examples from the Hadrien Gourlé's bioinformatics tutorials with the University of Agricultural Sciences, Sweden.**
- **“Process locally, backup remotely”, “Mind your nanoseconds”, are very important in bioinformatics!**



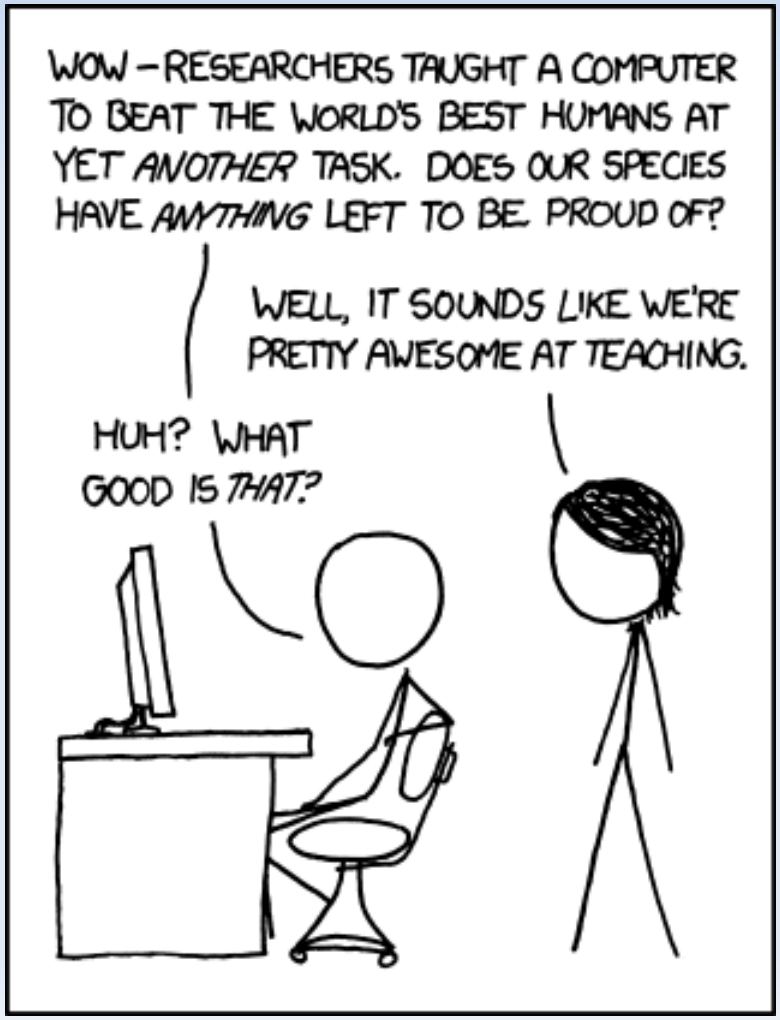
Conclusions

- **The provision of HPC systems correlates with research output (Apon et al, 2010). HPC generates excellent return-on-investment (Joseph, et al, 2013); about $44 in new income or cost-savings per $1 invested (almost entirely as positive externalities).**
- **VPAC study from change of cluster, usage hours, and divergence in training hours.**

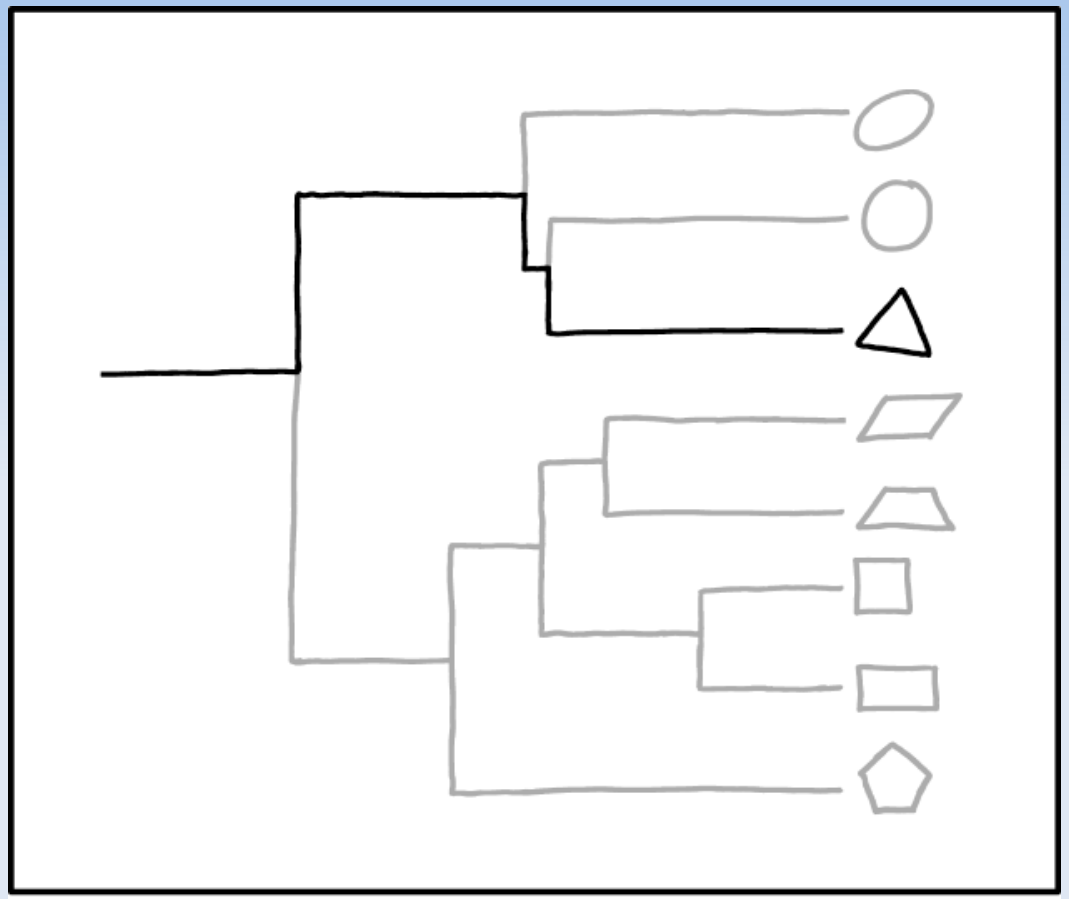
Year	RMIT	La Trobe
2012	1,729,837h	1,719,554h Tango
2013	8,108,695h	3,301,052h Trifid
2014	9,760,919h	4,964,297h Trifid

Trifid RMIT enrolments 229 La Trobe enrolments 29

- **At UoM in 2023, at least 54.14% of cluster utilisation measured by job submission was conducted by users after receiving training.**
- **Get the training right! Use people who are qualified and experienced in advanced adult education and HPC (no, there's not many!)**
- **Institutions that do not invest in HPC and user education have *will not survive*.**



Geometry and Genetics



THE PHYLOGENETIC REVOLUTION CONTINUES:

TRIANGLES WERE LONG BELIEVED TO BE RELATED TO SQUARES, BUT GENETIC ANALYSIS PROVES THAT THEY ARE ACTUALLY VERY POINTY CIRCLES.


THANKS FOR WATCHING
& LISTENING PATIENTLY
makeameme.org