Understanding the eResearch Ecosystem in New Zealand

Workshop Reflection Report

eResearch NZ 2018 Conference
Queenstown, New Zealand

Reflection Report for eScience Futures Workshop

written by Mark Dietrich and Nick Jones

August, 2018

Foreword

We started our journey as NeSI in mid 2011, and after 7 years we believe we've come a long way. As we look out another 7 years we recognise a changing landscape. Our NeSI Futures work intends to define the NeSI we'll need for our future, while retaining and building on our successes so far.

We're looking in three directions as we contemplate our future:

- Locally, to surface and connect with researchers and their research goals - where do they aspire to be in 7 years, as individuals and communities, and what role do they foresee for advanced research computing in getting them there.
- Internationally, as research exists in a global context, and the investments most similar to NeSI are those made by other nations within their own research systems - what are their strategies, where are they going, and what have they learnt along the way.
- Within our own organisation and across our network of collaborators, to appreciate the capabilities we already have, and those we aspire to - what should we focus on, what can we build upon, how do we partner, and where can we go, together.

Across these discussions we've identified a need to establish a common language. This initial report from our NeSI Futures workshop provides this foundation. We are using it in our discussions with international colleagues who are also subject matter experts in advanced research computing and eScience. It will be of most use to those inside the core business of eScience.

For those interested in our discussions with local researchers on their research goals, we'll publish additional findings across 2018, so look out for these over time.

I've worked closely with my colleague Mark Dietrich in shaping up the workshop reported here, and on the ongoing work we're doing to canvas colleagues internationally. Mark comes from a recent role of running Compute Canada, a leading international eScience investment with many similarities to our own NeSI, though obviously one of those is not scale! Mark's experiences in leading Compute Canada through a period of significant evolution and development and the insights he's gained provided common ground when we met up at SuperComputing 2017 in Denver. It was there we first formulated a plan to conduct an international benchmarking study, and from there that this workshop took shape.

As we publish this initial report, our work has moved on, and we're nearing the end point of our international benchmarking. We're well progressed on our local researcher consultations. And we're heading into our first stakeholder workshops to explore our own next steps. We'll follow this publication with another, looking at what we've learnt internationally. We'll also share insights gained from researcher discussions before the year is out.

Nick Jones
August, 2018

Prologue

eResearch, Advanced Research Computing, Digital Research Infrastructure, eInfrastructure, Cyberinfrastructure – there is little agreement about what to call it, but there is growing agreement that, whatever you call it, it underpins, enables and is increasingly essential to the most significant research around the globe. Discovering the *God particle* (the Higgs Boson), proving the theory that gravity waves exist, personalizing medical treatment based on our DNA, designing drugs, materials and products that can change our world – the list of societal, scientific and engineering achievements that would not be possible without eResearch gets longer and longer.

Given the importance of eResearch, it is surprising that there has been little attention paid to *what* constitutes an eResearch service and *how* each service can best be provided to researchers. The cost of these services, particularly the capital and operating costs associated with leading high performance computing systems, are significant and almost universally covered by the public in one way or another, so it is important to make smart decisions about what services to provide and how to provide them. This report represents the start of such an investigation, into both the *what* and the *how* of eResearch. Later reports will advance the investigation, defining the components of eResearch ecosystems found around the world, describing how those ecosystems deliver those services, and identifying the ecosystem characteristics that are most effective in enabling their "client" research communities to compete on the world stage.

This last point, identifying the characteristics of effective eResearch ecosystems, will help NeSI chart its own course for the next 5-7 years and help position NeSI to be the effective eResearch partner that New Zealand researchers need to continue to compete and excel globally. Of course, New Zealand is not alone in this examination – research roadmapping efforts are the norm in many countries, initially motivated by the investment magnitude and long lifetimes of many research infrastructures, such as the Large Hadron Collider at CERN in Switzerland, which is now about to celebrate its 10th anniversary of operations. Australia just completed a national roadmapping exercise, the UK is in the midst of a comprehensive effort, and the US is plotting its cyberinfrastructure needs through 2030. As the Red Queen says in *Alice in Wonderland "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"*

Since eResearch investments suffer not from long, but extremely short, useful lives, learning how to design a more sustainable, agile and effective eResearch ecosystem might create more of a competitive advantage for New Zealand than chasing technical measures like cores, flops or bytes. Not only will such sustainability and agility serve New Zealand research well, it will also serve as an incubator for solutions and technologies that could themselves compete on the world stage, as well as developing the skilled people who are in the greatest demand from leading enterprises around the globe. As this study unfolds, we hope to find the key factors that will unlock such a bold future for New Zealand.

Mark Dietrich
August, 2018

Acknowledgements

The authors acknowledge and thank:

- The workshop attendees for contributing in the *eScience futures workshop* - see **Appendix C** for the full list of attendees
- NeSI Board members and MBIE colleagues for consultation and contribution throughout the shaping of this document
- Jun Huh and Jana Makar for helping with content revision and formatting of this document

Preface

NeSI has commissioned a study of comparator jurisdictions and facilities from around the world, to inform NeSI's proposal with relevant examples from those international benchmarks. This study will capture major international trends in advanced computing technology and digital science, as well as the evolution of relevant services available from a range of providers. In parallel, NeSI is running research community consultations, reaching out to NZ research leaders in order to capture a broader understanding of research drivers in the timeframe of 5 to 7 years in the future. These activities began in late 2017 and should be completed by late 2018.

eResearch New Zealand is an event held early each year that assembles experts with an interest in eResearch from across NZ and across a range of research disciplines. In 2018 this event coincided with the start of the studies noted above, so several workshops were included in the event program to allow the eResearch community to participate in the early stages of the needs assessment and ecosystem benchmarking activities.

What does our eResearch ecosystem look like?

Facilitated at the 9th annual eResearch New Zealand conference, one of the workshops hosted was *eScience futures workshop*, to build shared language and common understanding of both local and international eScience value chains in the context of global developments. This document is a reflection report of this workshop.

This report aims to provide a springboard for further discussions with NZ research communities and with representatives from key comparator facilities and jurisdictions around the world, as well as for validation of key conclusions with a number of experts from around the world. This feeds into broader discussion NeSI is facilitating on relevant investments required to underpin and enable New Zealand's longer term research directions, primarily to inform NeSI's future business case from mid-2019.

Audiences for this document include:

- NZ research communities. NeSI will use this document to set the stage for assessing eResearch needs for 2025.
- Representatives of international comparator facilities and jurisdictions, to inform international benchmarking.
- eResearch experts, to validate value chain elements, maps, and key assumptions and build shared language.

Appendix A: Overview of New Zealand Research *describes a NeSI-specific view of the NZ situation for audiences beyond NZ.*

Going Forward

NeSI will reach the end of its most recent funding mandate next year (2019) and is developing a detailed business case for a follow-on funding proposal to government. NeSI's case for its most recent capital investments[1] contemplated the introduction of several new services and capabilities based on a National Platforms Framework Review,[2] its assessment of requirements from the NZ science community. At the same time, the evaluation of NeSI's performance since 2014 posed questions about the extent of such requirements and whether NeSI was the best organisation to deliver some of those services if confirmed to be required. At a higher level, that performance evaluation also encouraged NeSI to consider the overall eScience ecosystem in New Zealand, as well as learnings from relevant comparator facilities and countries around the world, before including specific services or capabilities in its next funding proposal.

[1] https://www.nesi.org.nz/services/high-performance-computing/platforms/national-platforms-framework-2015-revision

[2] https://www.nesi.org.nz/news/2016/05/nesi%E2%80%99s-national-platforms-framework-consolidation-refresh

Introduction

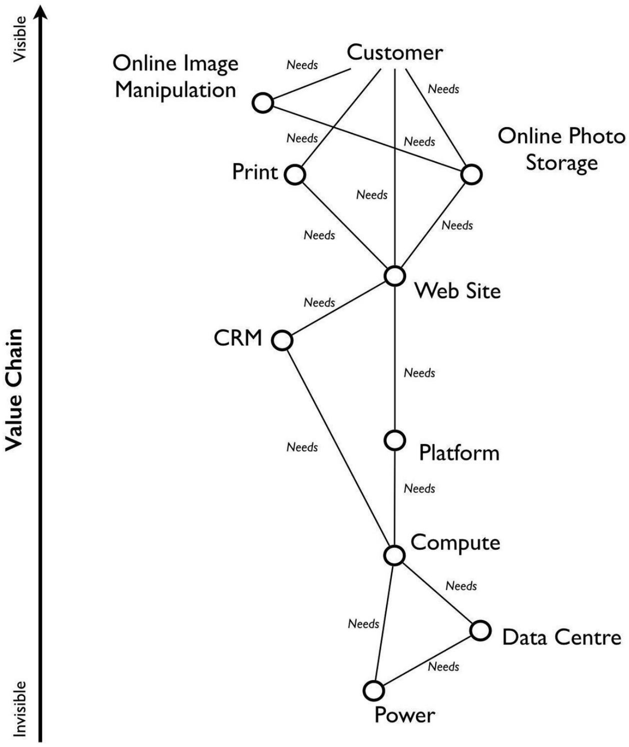
NeSI's preparatory activities must focus on the needs of the New Zealand research community, and the intent is to benchmark them against facilities and jurisdictions around the world, so it is important to create a globally consistent framework to enable such comparisons. The needs of specific researchers from different disciplines, even within New Zealand, also vary widely, so this framework must capture requirements from multiple disciplines with a diverse range of needs.

eScience Futures workshop has introduced value chain mapping to the participants, establishing a preliminary framework for understanding the eResearch value chain. Participants divided into break-out groups to drill down into several high level components to discuss subcomponents, refine their definitions and build shared understanding. The objective was to create a common foundation for the community discussions that will follow.

Introducing Value Chain Maps

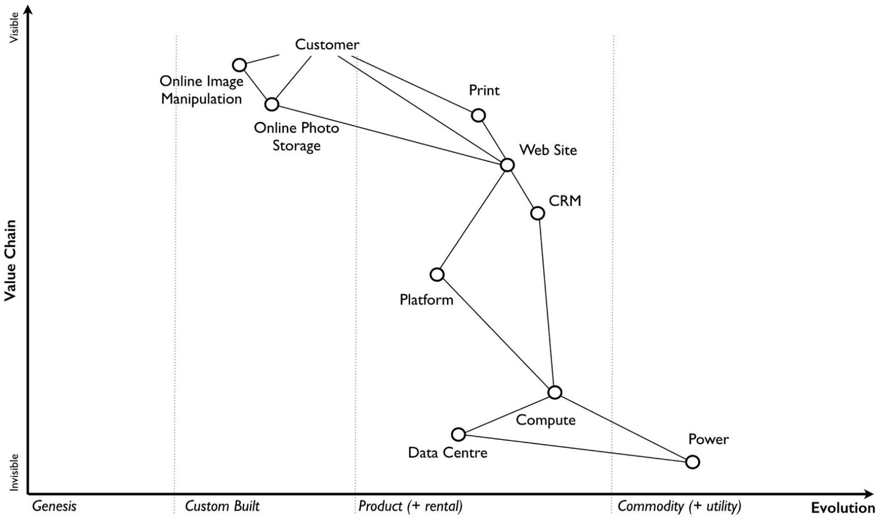
To create a common basis and shared language for discussion, NeSI is using a technique called *value chain mapping*, originally created by Simon Wardley, to identify and categorise the many activities that different research groups associate with eScience.

Wardley's technique is often used to graphically show how a user's or customer's specific need, indicated at the top of the value chain map, is met through combinations of services, or *components*, each of which in turn requires its own components, forming a cascading chain of components that range from *advanced* or *high level* at the top to more fundamental as you move to the bottom of the map. Wardley's first map, of an online photo service he was running, is shown at the right.



Value chain maps allow the *evolution* or *maturity* of each component to be recorded graphically along a horizontal axis, showing whether the component is a commodity (at the far right of the scale) or an early stage experimental component (at the far left). Experimental components are positioned at the left of the map, while commodity components are placed at the far right, with *custom* products or services falling somewhere in the middle. There are no hard rules about this placement, but those familiar with the components can usually agree on relative placement. Reaching agreement on placement also typically requires achieving rough agreement on what each component actually entails, i.e. on its definition.

Here is the value chain again, redrawn to also show the stage of evolution of each component:



Given the rapidly evolving nature of eScience and the role that experimental services frequently play, the ability to display this aspect of the value chain components is very useful.

For additional background on Simon Wardley's technique, visit: https://medium.com/wardleymaps/on-being-lost-2ef5f05eb1ec

Workshop:

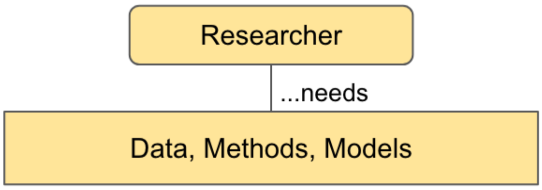
Building an Understanding of the eResearch Value Chain - and the eResearch Ecosystem

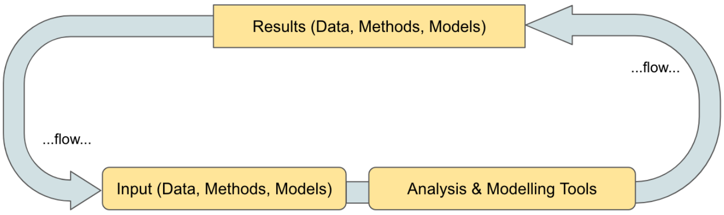
NeSI's preparatory activities must focus on the needs of the New Zealand research community, and the intent is to benchmark them against facilities and jurisdictions around the world, so it is important to create a globally consistent framework to enable such comparisons. The needs of specific researchers from different disciplines, even within New Zealand, also vary widely, so this framework must capture requirements from multiple disciplines with a diverse range of needs.

This workshop introduced value chain mapping to the participants, establishing a preliminary framework for understanding the eResearch value chain. Participants divided into break-out groups to drill down into several high level components to discuss subcomponents, refine their definitions and build shared understanding. The objective was to create a common foundation for the community discussions that will follow.

Preliminary Framework for the eResearch Value Chain

Value chain mapping starts with the needs of the user, shown at the top of every value chain. Ultimately, and very simplistically, all a researcher needs from the eResearch value chain are *results*, in the form of research data, methods or models.

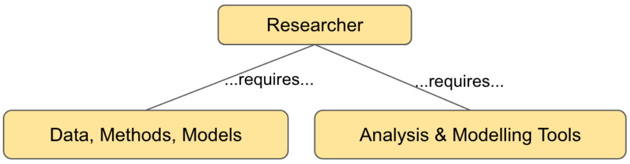


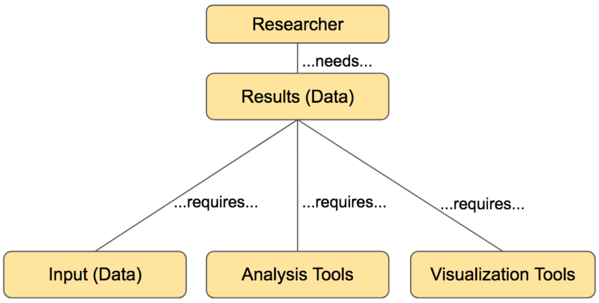

Results (Data, Methods, Models)
...flow...
...flow...
Input (Data, Methods, Models)
Analysis & Modelling Tools

Research results flow from inputs (again, in the form of data, methods and models) combined with analysis and modelling tools.

Ideally the results of one research project become inputs to new research projects.

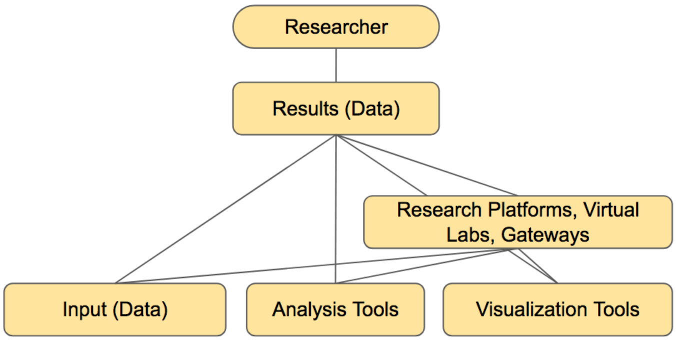
However, value chains do not illustrate the flow of elements, activities or functions. Instead they illustrate the lower level *components* required by each level above.

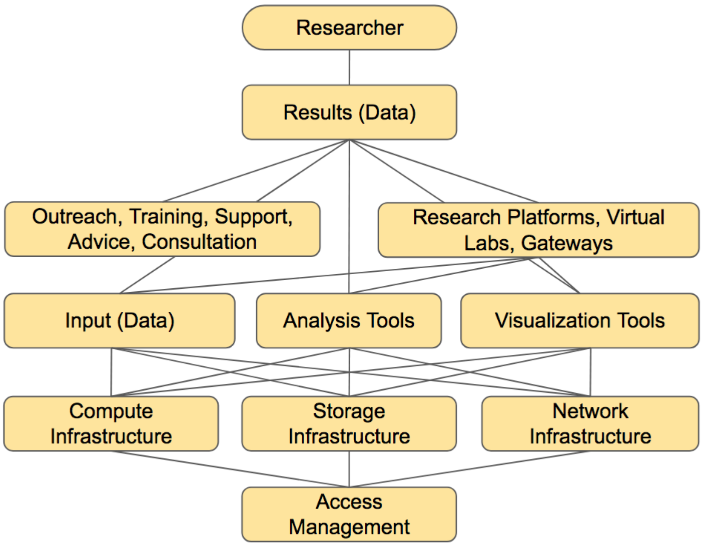




Armed with knowledge about the eResearch components needed, researchers can use them in different combinations to generate the results they need:

Since different researchers might follow different sequences of activities, some researchers need tools to help them manage this process, and these tools fall into a general category of *research platforms, virtual labs and science gateways*, mapped here.



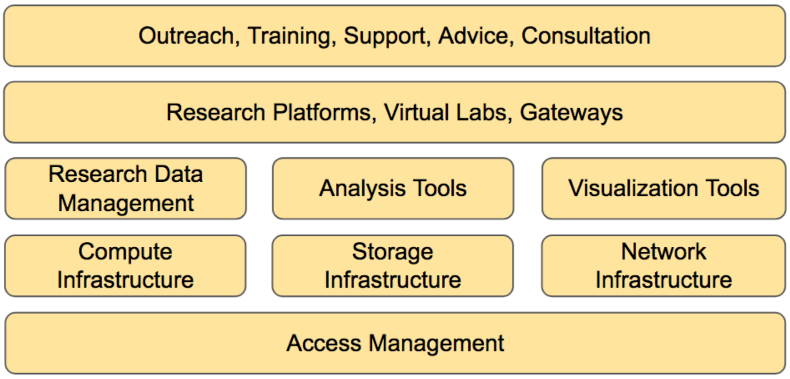


Researchers combine different eResearch services (components of the eResearch value chain), depending on the results desired and the research being conducted. Some researchers know the services they need, but every researcher has to learn this to begin with, so some of the first services a researcher needs are outreach, training and support, consultation and advice, and perhaps even collaboration with data or computational scientists. For now we group these services together in a single component.

When we consider the *input data*, *analysis tools* and *visualisation tools* components, they in turn rely on more traditional compute, storage and network infrastructure components, again in varying combinations. Finally, all of these components need mechanisms to manage access by, and authentication of, users. Assembling all of the components in the map, we arrive at a fairly complete but high level map of the full eResearch value chain.

Even at a high level, this map is complex and has many interconnected components. To make subsequent maps easier to read, we hide the lines that symbolise how one component requires a component at a lower level and simply assume that any component at one level might *require* components at lower levels. This allows us to simplify the map.

This sort of diagram is similar to a traditional technology stack diagram, but as discussed above the vertical dimension has a very specific meaning in a value chain map.



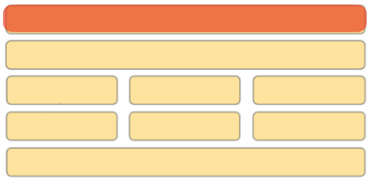
In the map above, in the two rows with multiple components, the components have been placed very roughly to reflect their comparative stages of evolution. For example, research data management has recently been recognised as a distinct activity, earning a place on the left, while many visualisation tools and techniques have been in use for decades (if not centuries!). Of course this is a very simplistic treatment, and a more detailed exploration of each of these components, identifying the embedded subcomponents, shows that even within a given component the maturity of subcomponents varies widely.

In the following sections, each of the components mapped above will be explored in more detail with key trends and value chains of the subcomponents at more granular level.

For all following value chain maps, each component has been *coded* in three colours to reflect discussions during and in parallel with the workshop, as well the ongoing refinement of the maps themselves:

- Yellow components were included in the preliminary value chains presented in the workshop.
- Orange components were identified by attendees as required.
- Green components represent important services and activities brought up in other presentations at the event.

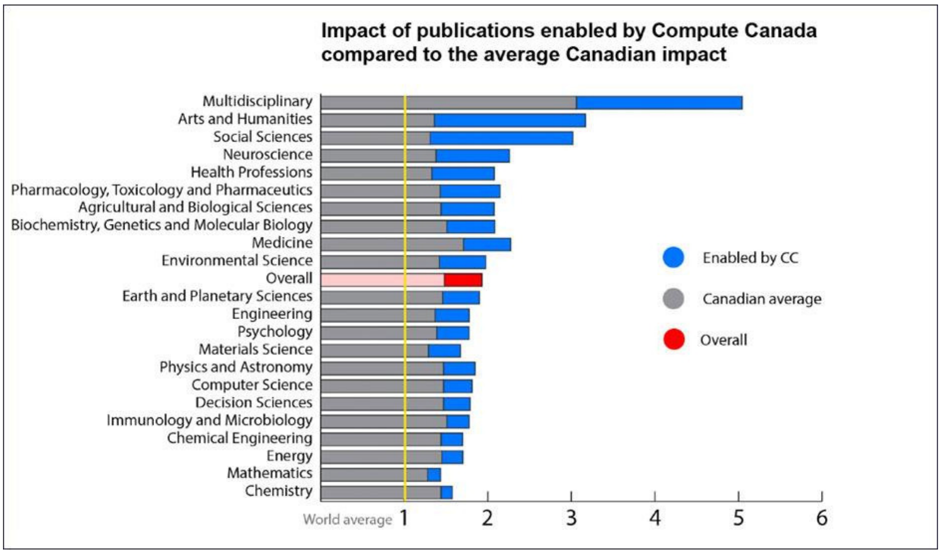
Outreach, Training, Support, Advice, Consultation



This high level component encompasses a wide range of personal interactions between users and a variety of experts associated with the eResearch facility and related services. Interactions can range from traditional training and outreach activities, to interactive support, advice and consultation activities. With the growing range of eResearch facilities and services, as well as the growing awareness of how research activities in all disciplines can be enhanced through the use of such capabilities, researchers often need human guidance to determine which capabilities to use and how to integrate them into their own activities.

Key Trends

Across multiple disciplines, the number of researchers interested in integrating eResearch capabilities into their own research is growing. In disciplines like chemistry and physics, where eResearch is an established tool, exploiting this tool becomes increasingly essential to globally competitive research. In other disciplines, such as the humanities, where eResearch is less-commonly used, adopting these techniques has been linked to increased research impact (diagram from Compute Canada bibliographic study), also prompting growing interest.



Growing interest creates growing demand for the range of activities included in this component, challenging the eResearch ecosystem to scale its outreach and training activities to respond to this growing demand. The range of possible topics, from learning the basics to applying more advanced techniques, also challenges the ecosystem to deliver the range of specialised services required. In addition, the growing importance of eResearch to research in general creates a need to build awareness and understanding with a broader range of audiences, including government and public stakeholders, individuals and groups affected by eResearch (for example as potential subjects), as well as actual and potential users.

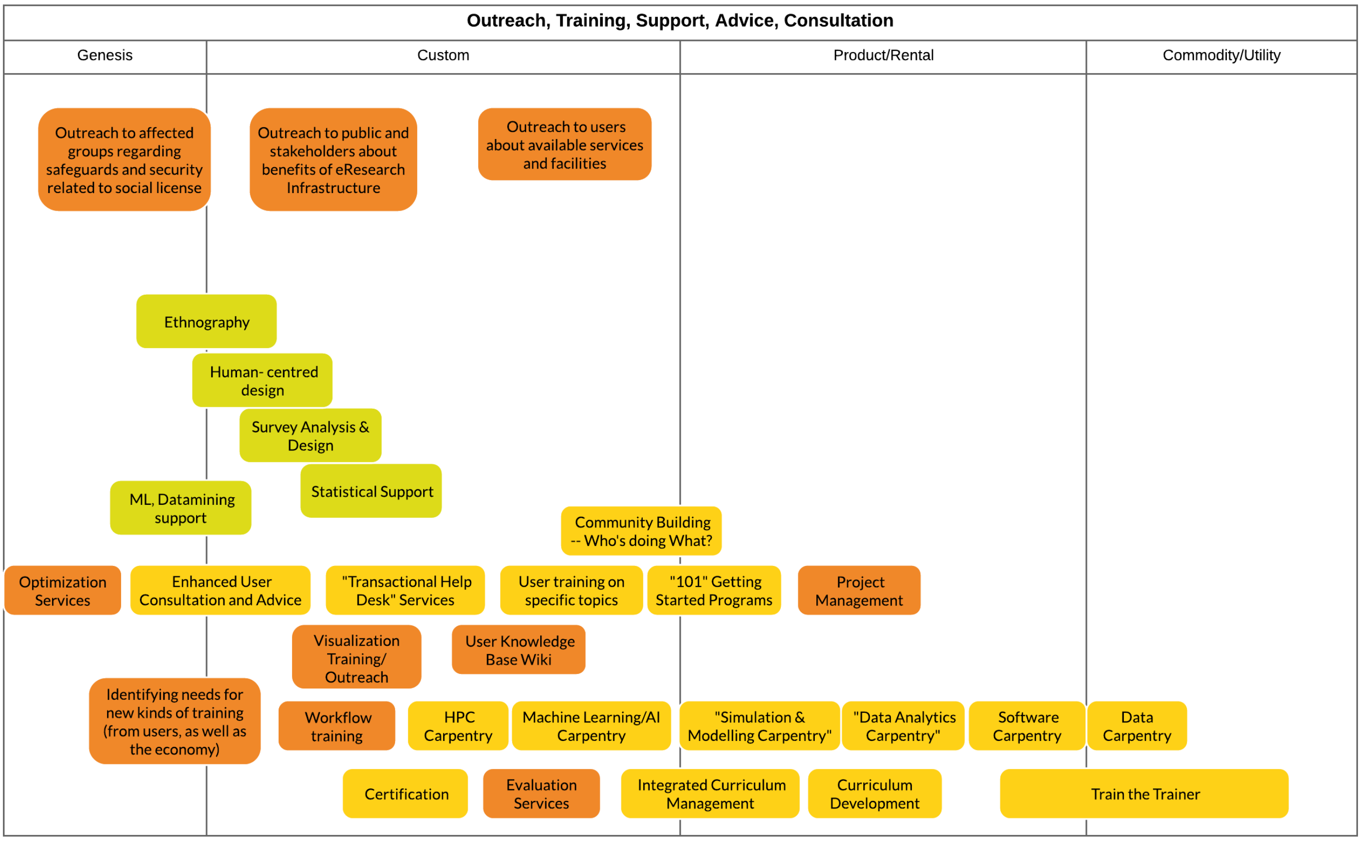
Historically eResearch ecosystems have responded to these challenges by dedicating more staff to person-to-person outreach, training and support, and creating classroom style teaching mechanisms (familiar to researchers and eResearch experts alike). However it is not clear if these approaches can scale efficiently to the levels needed. Ideally research groups would like to have appropriately trained resources, for example in research data management, *embedded* within their research teams, implying that eResearch techniques need to be more commonly known among the research community, rather than *bolted on* with just-in-time training.

These types of challenge are not unique to eResearch. Reaching out to groups of users and potential users, building awareness of available facilities and services, leading users on a journey from novice to expert, creating an *error-free* user experience -- in the business world, these activities are known as *marketing*, *sales* and *support*, and there are many ways to meet these challenges effectively, while at the same time maintaining high levels of user satisfaction.

Current responses to growing levels of demand include the *carpentries* -- such as software carpentry, which is essentially a standard course curriculum distributed worldwide through a *train the trainer* program, as well as the creation of standard curriculum guides, such as those distributed by HPC University (http://hpcuniversity.org/). Even researchers familiar with eResearch techniques acknowledge gaps in basic topics such as Fortran, Python, Github and code optimisation, so standard curricula and carpentry-style courses are good starting points.

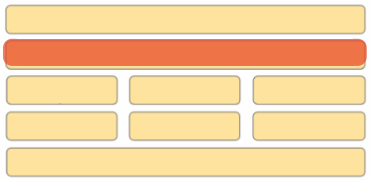
Both of these mechanisms give eResearch facilities the ability to serve more users more effectively with a limited number of staff. By contrast, the growing need for specialised support will create demand for robust training materials that can be easily customised for specific audiences (e.g. basic HPC skills for bioinformatics researchers) and will drive specialised communities to provide more support inside their own communities (e.g. community-based collaboration tools plus mechanisms that encourage support from within the community).

Value Chain of Subcomponents of Outreach, Training, Support, Advice, Consultation


Outreach, Training, Support, Advice, Consultation
Genesis
Custom
Product/Rental
Commodity/Utility
Outreach to affected groups regarding safeguards and security related to social license
Outreach to public and stakeholders about benefits of eResearch Infrastructure
Outreach to users about available services and facilities
Ethnography
Human- centred design
Survey Analysis & Design
Statistical Support
ML, Datamining support
Community Building -- Who's doing What?
Optimization Services
Enhanced User Consultation and Advice
"Transactional Help Desk" Services
User training on specific topics
"101" Getting Started Programs
Project Management
Visualization Training/ Outreach
User Knowledge Base Wiki
Identifying needs for new kinds of training (from users, as well as the economy)
Workflow training
HPC Carpentry
Machine Learning/AI Carpentry
"Simulation & Modelling Carpentry"
"Data Analytics Carpentry"
Software Carpentry
Data Carpentry
Certification
Evaluation Services
Integrated Curriculum Management
Curriculum Development
Train the Trainer

*See **Appendix B** for the full page view along with subcomponent definitions*

Platforms, Virtual Labs, Gateways



Broadly speaking, all of these components provide users with interfaces (typically graphically, and often web-, based) to a variety of applications, data and electronic services, making it easier for the user to perform a series of eResearch activities. Historically, eResearch capabilities and services required users to log in to a server with a username and password, and then type in properly formatted computer *commands* in response to a *prompt character* (such as >) in the server's user interface. Compared with today's visually-rich computer interfaces, such command line interfaces (CLI) are perceived by some as more difficult to use and more prone to error, creating barriers to adoption. Platforms, virtual labs, gateways, etc., provide more user-friendly interfaces to valuable services, replacing the command-line interface with the graphical interfaces with which many computer users are now familiar.

At a high level, this component also covers the various mechanisms that give users access to specific resources in an eResearch facility, resources such as compute time or data storage space, either through automatic, competitive or other means.

Key Trends

Easy to use is often requested by users, and significant investment has been made in the various intermediary services included in this component, with the objective of improving research productivity. Many *point solutions* such as community management tools, data sharing tools, electronic lab notebooks and workflow tools, have merged into integrated solutions, typically with different platforms focussed on the needs of different research communities, even though such platforms often deliver common functions. Some platforms have become de-facto standards for research in certain disciplines. Other platforms provide a common language for generating, documenting and sharing research results, such as *pipelines* used in certain bioinformatics research. With the growing emphasis on *reproducibility*, some platforms are adding tools to improve workflow and document processes, as well as capturing key metadata about the research data being used and produced as well as the analysis software and underlying computational resources being used to produce new results.

Established research platforms, science gateways, etc. are becoming increasingly difficult to maintain and keep relevant -- investments in those platforms have created *technical debts* that are expensive to maintain. Adding the functions mentioned above can be challenging for older platforms – leading to unstable or unreliable performance, or prompting expensive *rewrites* of the underlying code.

Rapid evolution of more generic (non-research focused) compute virtualisation (virtual machines, cloud-based services, containerisation, orchestration, microservices, lambda functions etc.) is making it possible for new eResearch platforms to be built quickly from generic platform components and services developed for other markets. Generic research platform frameworks, such as Apache Airavata, may enable rapid re-engineering of established platforms and lower the cost of on-going maintenance.

Mechanisms for accessing resources are also evolving. Historically the key resources were compute time and storage, but evolution of compute and storage infrastructure (as well as network infrastructure) is creating a wider range of *allocatable* resources:

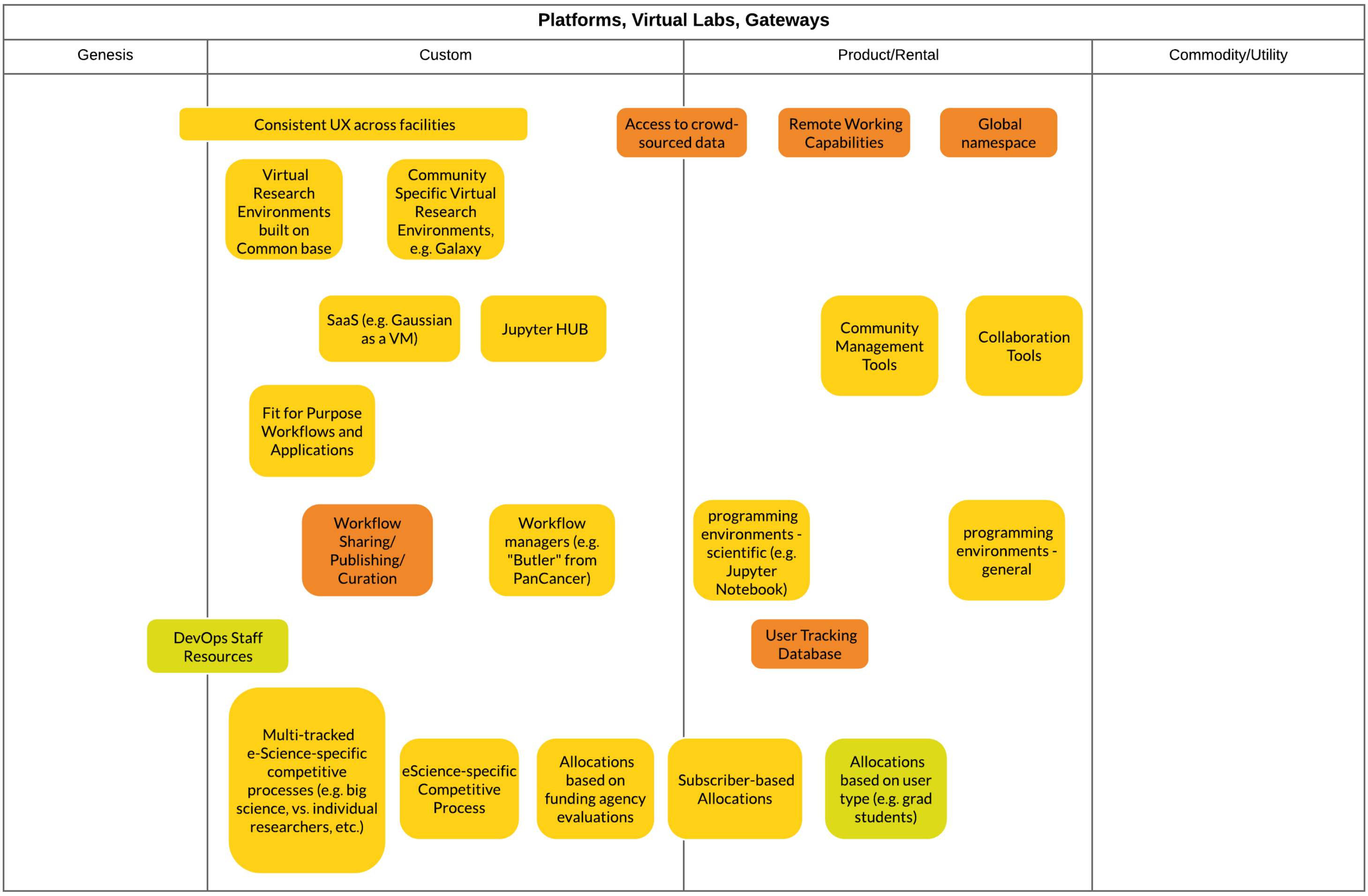
- Different kinds and capabilities of compute resources: new/old CPUs, accelerators, local memory, local disk, interconnect topologies, virtual machines, web servers, database servers.
- Different kinds and capabilities of storage resources: scratch storage, campaign/project storage, database capabilities, shareable research data, archiving/preservation/backup.
- New network infrastructure resources: permanent IP addresses, science DMZ access and authorisation.

Allocation mechanisms can include a range of factors:

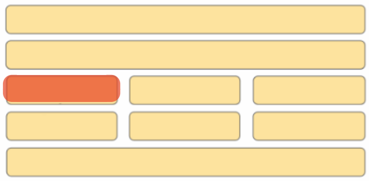
- The strategic, or mission-based, nature of the research, e.g. cancer genetics vs. archaeology, industry collaboration or discovery research, part of an international project (e.g. ATLAS), part of an ongoing program (e.g. weather forecasting, seismic analysis).
- Excellence of the research, perhaps based on peer-review of the related funding proposal, or separate peer-review of the specific proposal to use of eResearch resources,
- Scale of request – requests for large amounts of resource (e.g. compute core hours or years) may require additional peer-review steps or strategic assessment,
- Type of user, e.g. graduate students and faculty might be allocated minimum amounts of resource automatically, with additional resources available through an application process.

All of these mechanisms might be implemented to varying degrees depending on the rights of the economic sponsors of a given eResearch facility. For example some capacity might be allocated based on a national peer-reviewed evaluation process, while other capacity might be distributed to all faculty of an institution that contributes financially to the facility, based on that institution's own procedures.

Value Chain of Subcomponents of Platforms, Virtual Labs, Gateways


Platforms, Virtual Labs, Gateways
Genesis
Custom
Product/Rental
Commodity/Utility
Consistent UX across facilities
Access to crowd-sourced data
Remote Working Capabilities
Global namespace
Virtual Research Environments built on Common base
Community Specific Virtual Research Environments, e.g. Galaxy
SaaS (e.g. Gaussian as a VM)
Jupyter HUB
Community Management Tools
Collaboration Tools
Fit for Purpose Workflows and Applications
Workflow Sharing/ Publishing/ Curation
Workflow managers (e.g. "Butler" from PanCancer)
programming environments - scientific (e.g. Jupyter Notebook)
programming environments - general
DevOps Staff Resources
User Tracking Database
Multi-tracked e-Science-specific competitive processes (e.g. big science, vs. individual researchers, etc.)
eScience-specific Competitive Process
Allocations based on funding agency evaluations
Subscriber-based Allocations
Allocations based on user type (e.g. grad students)

Research Data Management



This high level component, a comparative newcomer to the eResearch landscape, provides the full range of services required to manage research data in order to make it *findable, accessible, interoperable, and reusable* (the *FAIR* principles of research data management), while at the same time preserving the data and protecting it from unauthorised use. In general, research data management sits at a relatively *exploratory* stage of evolution, with new tools, approaches and norms being developed all the time.

Key Trends

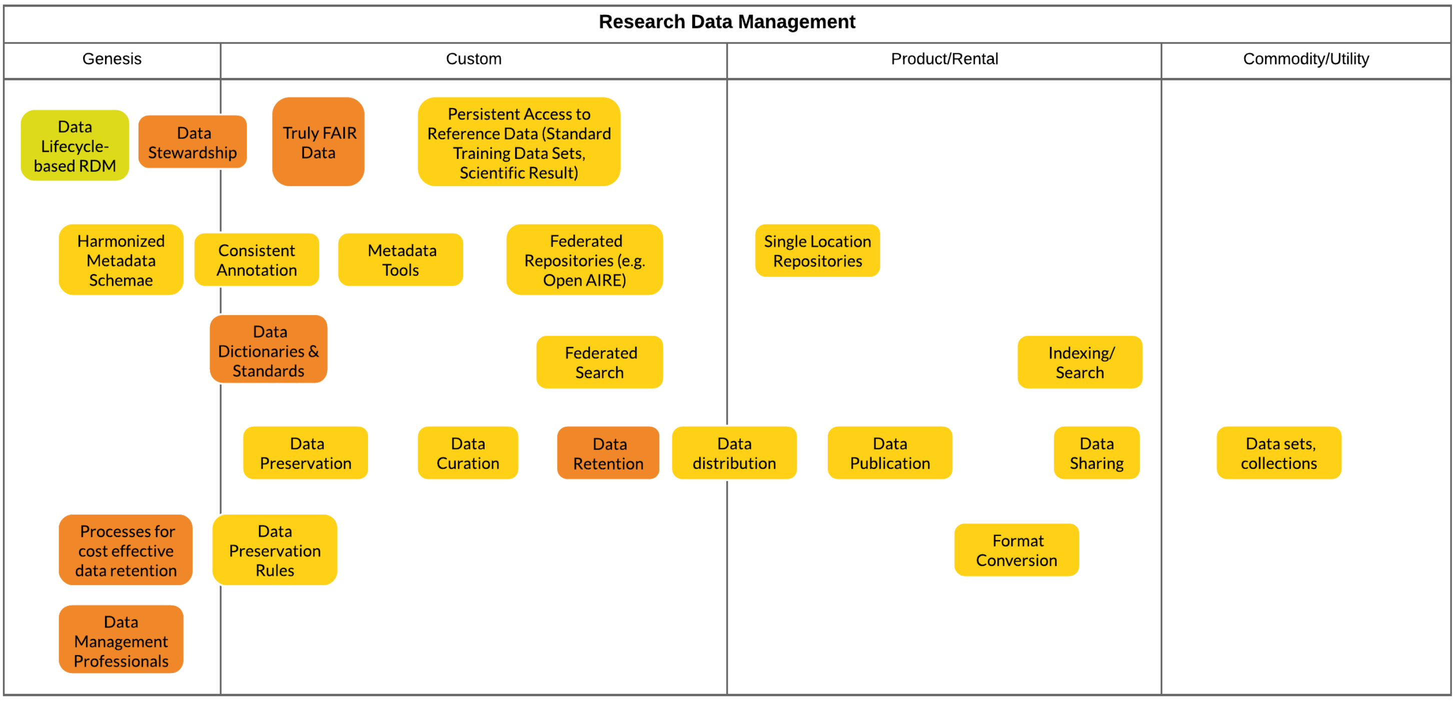
Research Data Management has been the subject of considerable effort at national, international and disciplinary levels. The common objective is to enable researchers to manage research data through its full life cycle (from creation to eventual possible destruction), and to make it *FAIR* as defined above. This contrasts with the sometimes poor practices of storing data on one or a few media, not considering the life expectancy of the media, not documenting the source, structure and possible uses of the data for anyone but themselves, and not making the data available to others or to the public. Worrying about longer term storage and preservation, documentation or sharing was the exception, rather than now, increasingly, the norm.

The new norm of preserved, documented and published or shared data triggers a range of requirements, from agreeing on metadata standards to creating economic and technical models for long term data preservation. Since all of the required components are *new*, aside from the underlying storage technology, all are undergoing considerable evolution and refinement, with advances in specifications as well as a variety of components being combined in specific service offerings.

Researchers themselves have a number of high level objectives, as reported during the workshop:

- Active, automated research data management
- Validation and provenance of research data
- Managing data produced by models, enabling use and combinations, providing catalogues
- Increased awareness of benefits of data sharing and publication. For example, cataloguing of prior results can avoid having to repeat time-consuming simulations.

Value Chain of Subcomponents of Research Data Management



Research Data Management

Genesis

Custom

Product/Rental

Commodity/Utility

Data Lifecycle-based RDM

Data Stewardship

Truly FAIR Data

Persistent Access to Reference Data (Standard Training Data Sets, Scientific Result)

Harmonized Metadata Schemae

Consistent Annotation

Metadata Tools

Federated Repositories (e.g. Open AIRE)

Single Location Repositories

Data Dictionaries & Standards

Federated Search

Indexing/ Search

Data Preservation

Data Curation

Data Retention

Data distribution

Data Publication

Data Sharing

Data sets, collections

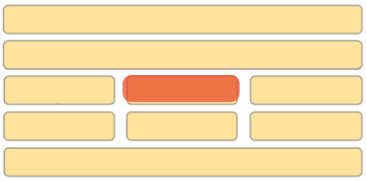
Processes for cost effective data retention

Data Preservation Rules

Format Conversion

Data Management Professionals

Analysis and Modelling Software



This component encompasses the many mechanisms for processing and analysing input data and/or for simulating or modelling the operation and behaviour of systems. Researchers increasingly have a range of analysis and calculation needs arising at different points in their research activities, ranging from pre-processing observational data to check quality and/or reduce data storage requirements, to *big data* analyses, to large scale simulations generating their own large output data sets, to post-processing of simulation output and machine-learning-based analyses of both observed and simulated data. Historically this has been the most visible component of the eResearch value chain, but for some disciplines, the analysis/simulation component, and the closely-related compute infrastructure component, play a smaller role than do storage infrastructure and visualisation.

Key Trends

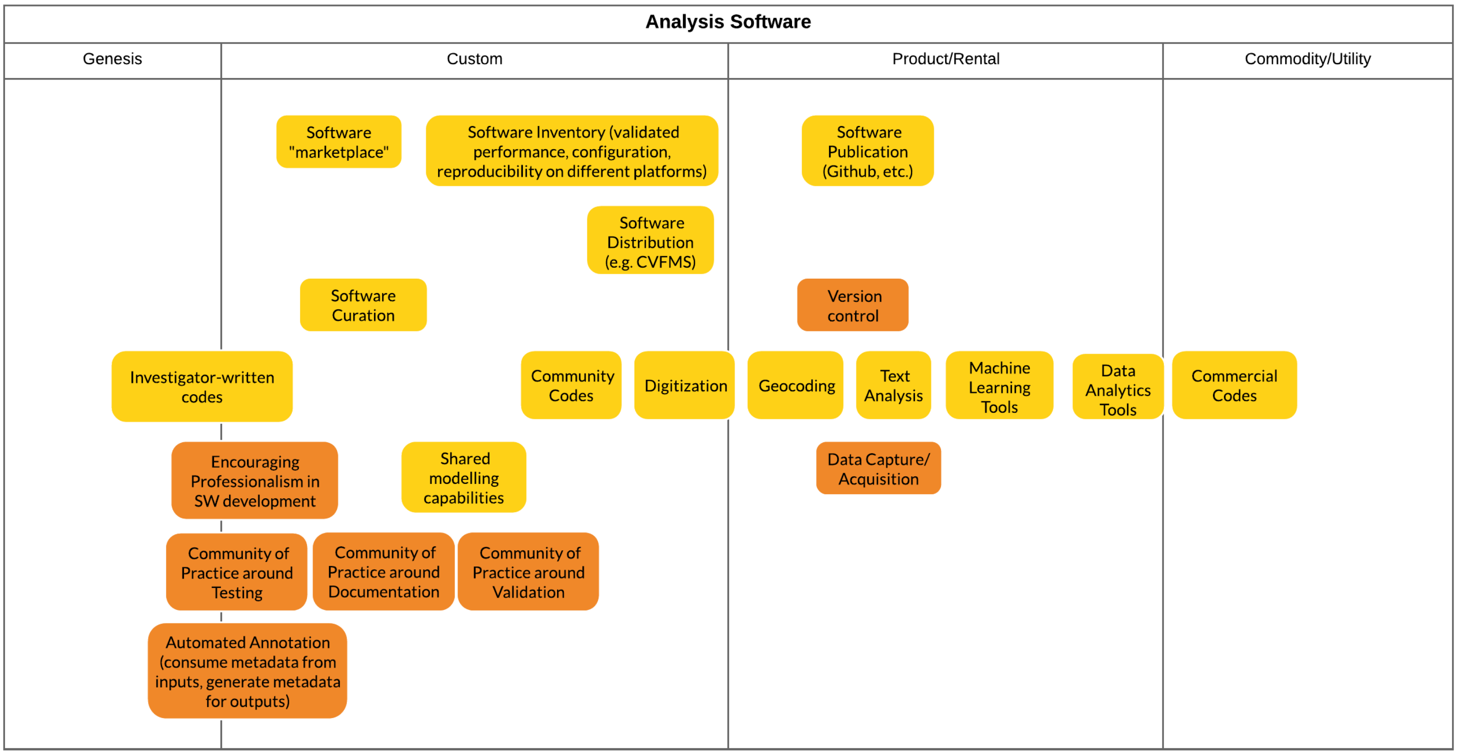
The overarching trend for this component is expanding scale: larger data sets, increasing scale of simulations (larger models and/or higher model resolution), and increased complexity and integration. Simulations are growing in size and accuracy to the point where their results can be aligned, compared and even tuned against experimental data; this is increasingly the norm in weather forecasting, and this approach is being applied in other fields as well. Biologists are looking at realistic ecosystem models and systems biology; earth and climate scientists are using more accurate models that incorporate more physical processes. Some key activities and models are gaining in refinement and power and becoming standards within specific communities. Some activities, such as genome assembly, are increasingly a commodity process, while other activities, such as data analysis (using either statistical or AI methods) would benefit from greater *packaging* to make them easier to use.

There is still a broad range of analysis/simulation tools available, and considerable effort is being devoted to creating new tools to meet specific requirements. This drives two key trends:

- Better software markets and inventories, married to effective software curation, allowing researchers to determine if quality tools already exist and are available for use.
- Improved *professionalism* among researchers who are developing new software or contributing to community efforts, addressing challenges in documentation, testing and validation.

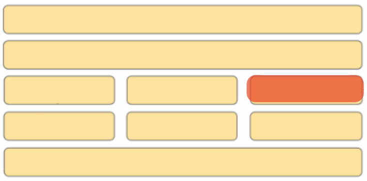
The first trend benefits from efforts in the area of research data management, since the software required to process and access research data is also an artifact that needs to be curated, managed and preserved.

Value Chain of Subcomponents of Analysis and Modelling Software


Analysis Software
Genesis
Custom
Product/Rental
Commodity/Utility
Software "marketplace"
Software Inventory (validated performance, configuration, reproducibility on different platforms)
Software Publication (Github, etc.)
Software Distribution (e.g. CVFMS)
Software Curation
Version control
Investigator-written codes
Community Codes
Digitization
Geocoding
Text Analysis
Machine Learning Tools
Data Analytics Tools
Commercial Codes
Encouraging Professionalism in SW development
Shared modelling capabilities
Data Capture/ Acquisition
Community of Practice around Testing
Community of Practice around Documentation
Community of Practice around Validation
Automated Annotation (consume metadata from inputs, generate metadata for outputs)

Visualisation

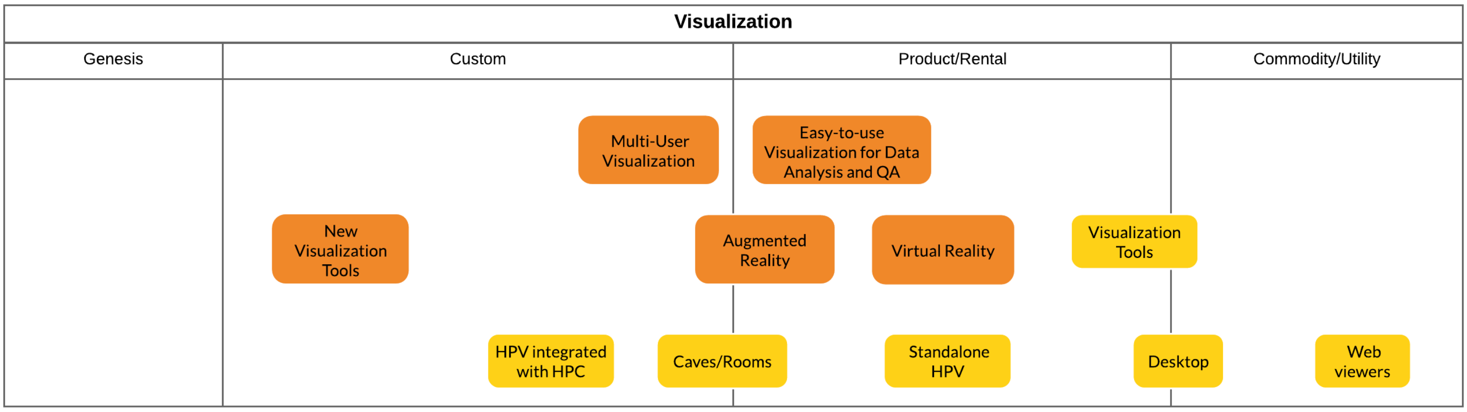
For centuries, scientists have presented data in a visual format to make it easier to find patterns and gain insight into otherwise hidden relationships in the data. Visualisation in the eResearch value chain encompasses the many tools used by different researchers to render data, sometimes large amounts of data, in a way that exposes relationships and hidden meaning.



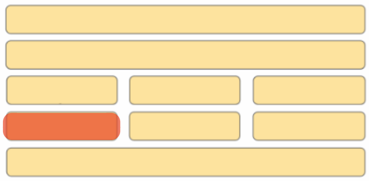
Key Trends

Major components of visualisation activities are data reduction and/or feature extraction, image/video rendering, and display and control (for example pan-tilt-zoom controls to change the point of view used for rendering). Domain-specific feature extraction algorithms (for example identifying *streamlines* in computational fluid dynamics (CFD) output results) are being generalised across domains, evolving to more general pattern-recognition functions and even transitioning to the use of machine learning and artificial intelligence. As output data sets from simulations expand in size, moving them to separate post-processing systems is increasingly difficult, so *high performance* feature extraction functions are increasingly executed on the data *in situ*, with smaller versions of the data sets moved to systems supporting visual exploration. Rendering performance and capability is increasingly impacted by developments in animation and film-making. Finally increased availability of high-resolution display devices (e.g. 4K video monitors and gaming goggles) makes previously expensive man-machine interfaces very affordable. Consumer gaming and related activities have also made drivers and interfaces for these devices very accessible.

Value Chain of Subcomponents of Visualisation



Compute Infrastructure



Along with the range of analysis needs seen by researchers (described earlier), many kinds of compute infrastructure are needed to efficiently support the different kinds of analyses and modelling required by researchers. Historically compute infrastructure has been the component of the eResearch value chain most people think of when they think of eResearch; in terms of financial requirements and funding, Compute Infrastructure also typically represents the majority of investment. Researchers concerned with the speed at which their analyses or simulations can run often demand the use of special kinds of compute infrastructure. Not all researchers need such cutting edge compute infrastructure, and instead can use commercially-available systems, as well as commercial data centre services now available in the cloud.

Key Trends

Price/performance ratios have continued to fall, consistent with Moore's Law, allowing the exponential growth in research computing requirements to be met with roughly constant levels of capital investment, assuming the right balance of compute, local storage, input/ output (I/O) bandwidth and high-speed interconnect performance can be found for any given analysis/modelling task or portfolio of tasks. Integration of graphical processing units (GPUs) into compute architectures, along with appropriate vectorisation/parallelisation of related analysis/modelling software, is contributing to increasing capabilities. Wider adoption of FPGAs, ASICs and tensor processing units (TPUs) have also enabled speedups for selected codes.

On premise installation of research compute infrastructure is increasingly being challenged by deployment *in the cloud* through virtualisation/orchestration technologies such as OpenStack, which allows on-demand access to portions of larger physical systems located in one or more remotely-located, enterprise-grade data centres. Most business IT requirements are well met with largely standard server designs, and these can be configured and added to, or dropped from, a virtual cluster to meet changing demand (e.g. from employees accessing enterprise applications as they arrive at the office, or shoppers accessing e-commerce applications driven by advertising). By contrast, the hardware requirements for research computing are diverse and specialised, particularly for large-scale simulations that require many tightly-interconnected processors not typically required in a business IT environment. However, as engineering, pharmaceutical and machine learning applications create incentives to add accelerators into cloud data centres, this is also becoming an alternative route for research computing. Even high core-count interconnected clusters are becoming available on demand, with Infiniband-connected instances of up to 128 cores available now on Amazon, and larger clusters available through services such as ReScale.

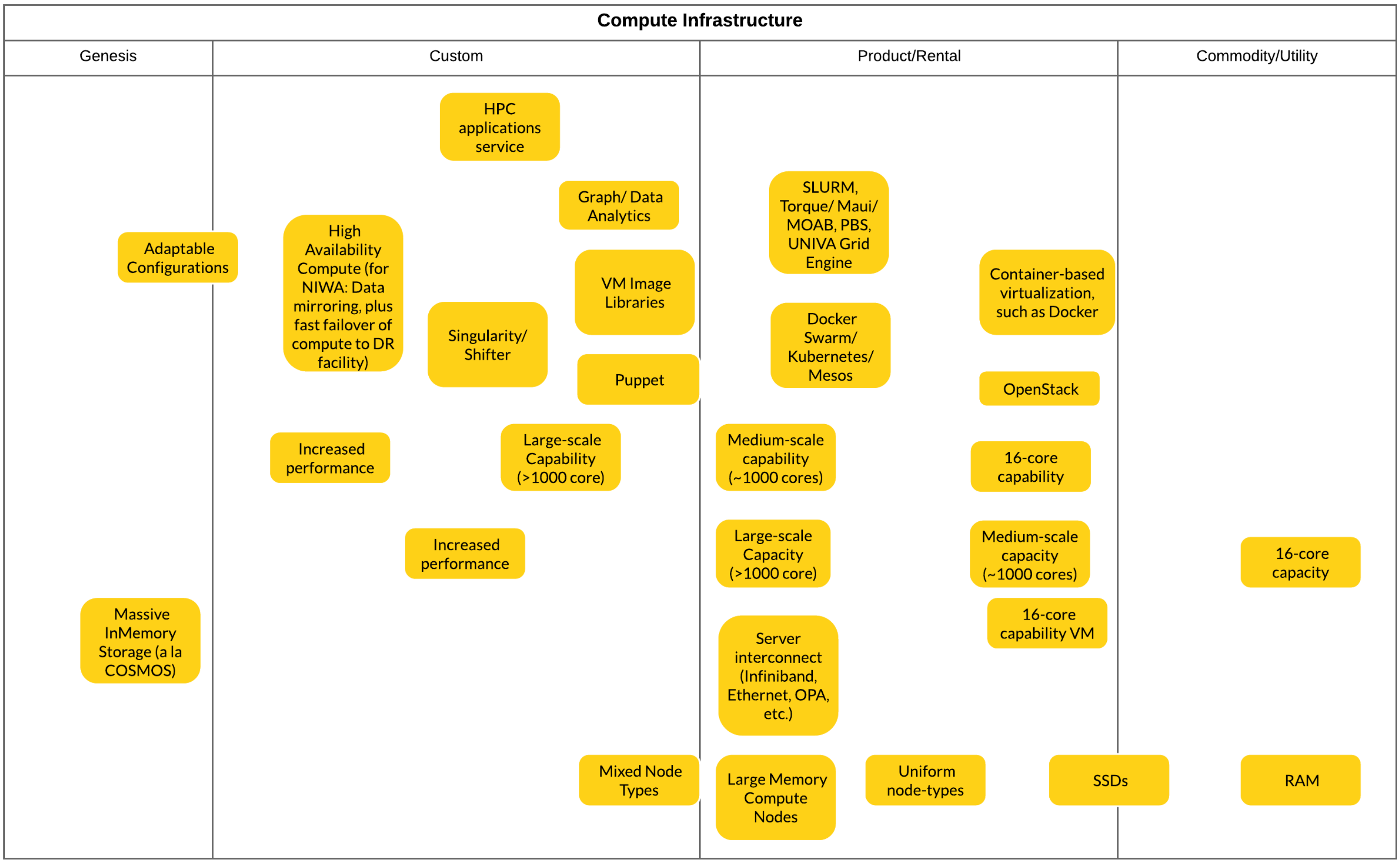
The challenge then becomes one of balancing architectural choices against procurement mechanism and costs, and against the costs/obstacles of bringing the data to the chosen compute infrastructure. *Renting* systems in the cloud provides some of the latest hardware to be accessed, but at up to 5x the total cost of ownership of a purchased system. Purchased systems have a fixed architecture (although they can be expanded through additional investment), so care must be taken to specify a mix of hardware that has the most general utility for the intended user base. Computations working with or producing large data sets require added time and in some cases costs for data movement, unless the data is stored near the compute resource. Not all analysis software can take full advantage of GPUs and similar augmentation technologies.

The predictability of a research computing requirement, as well as the magnitude of the data involved, might be the biggest factors determining whether on premise or cloud solutions are best. Weather forecasting, earth science and long-lifetime high energy physics experiments (e.g. ATLAS) are excellent candidates for on premise solutions. There is also great potential for optimisation of bioinformatics applications, given the large and growing data sets involved, but both the analysis tools and the hardware required continue to evolve, making it hard to choose an architecture that will last 5 years.

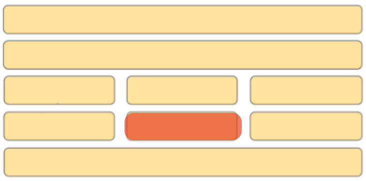
The processing of large data sets by analysis and modelling tools requires rapid access to specific data records and sometimes high speed parallel transfers of those records to the compute system, increasingly creating a performance bottleneck for certain analyses. For this reason, high performance *scratch* storage tightly connected to compute systems is typically considered as part of compute infrastructure configuration and design, rather than storage infrastructure, in order to optimise performance. Alternatively, data can be loaded into random-access memory configured as part of each computer processor, provided there is enough RAM available in the system. These considerations of file I/O speeds and local memory capacity strongly affect system design and performance running certain kinds of analysis and modelling tools. They also can reduce or eliminate cloud-based compute as an option for some analysis tasks if the required storage performance or memory capacity are not available.

As a final consideration, having local access to large compute resources is seen as a *reputational* factor that can affect the ability of researchers to participate in international projects and grand challenge initiatives. Care must be taken to separate procurement choices from questions of research strategy and competitive aspiration.

Value Chain of Subcomponents of Compute Infrastructure


Compute Infrastructure
Genesis
Custom
Product/Rental
Commodity/Utility
HPC applications service
Graph/ Data Analytics
SLURM, Torque/ Maui/ MOAB, PBS, UNIVA Grid Engine
Adaptable Configurations
High Availability Compute (for NIWA: Data mirroring, plus fast failover of compute to DR facility)
VM Image Libraries
Container-based virtualization, such as Docker
Singularity/ Shifter
Docker Swarm/ Kubernetes/ Mesos
Puppet
OpenStack
Increased performance
Large-scale Capability (>1000 core)
Medium-scale capability (~1000 cores)
16-core capability
Increased performance
Large-scale Capacity (>1000 core)
Medium-scale capacity (~1000 cores)
16-core capacity
Massive InMemory Storage (a la COSMOS)
16-core capability VM
Server interconnect (Infiniband, Ethernet, OPA, etc.)
Mixed Node Types
Large Memory Compute Nodes
Uniform node-types
SSDs
RAM

Storage Infrastructure



With the explosion of *big data* in all walks of life including research, the scale of storage infrastructure required to store and preserve valuable research data is increasing at an exponential rate, to the point where investment in storage infrastructure must grow to a scale comparable to compute investment. This in turn is driving evolution in storage infrastructure from *one size fits all* to a variety of storage capabilities adapted to a range of research data storage needs.

Key Trends

As with compute infrastructure, price/performance continues to fall predictably for storage systems, although at a lower rate than for compute (roughly 17% per year per *byte* vs. 40% per year per compute *flop*). Unlike compute infrastructure – with many types of compute infrastructure, there are essentially two *types* of storage infrastructure – disk storage appropriate for longer term storage of data, and tape storage, with lower costs and still lower transfer speeds appropriate for data backup and archiving. (As noted above, high performance *scratch* storage tightly connected to compute systems is treated in these value chain maps as an integral part of compute infrastructure.) In most implementations, basic storage capacity is augmented or packaged into storage services using a variety of storage management software, enabling features such as object storage, hierarchical storage management, georeplication and backup, metadata storage, etc.

As with compute, cloud-based storage solutions are available – and as with compute, cloud-based storage currently costs significantly more than the total cost of ownership of on premise storage capacity. Cloud-based storage is usually accessed as a storage service, offering many of the features that would require installation and operation of storage management software in an on-premise implementation.   Cloud-based storage does not offer physical tape backup capabilities, so facilities looking for physical tape copies of files in order to play a role in long term data backup, archiving or preservation will need to establish on premise tape backup systems.

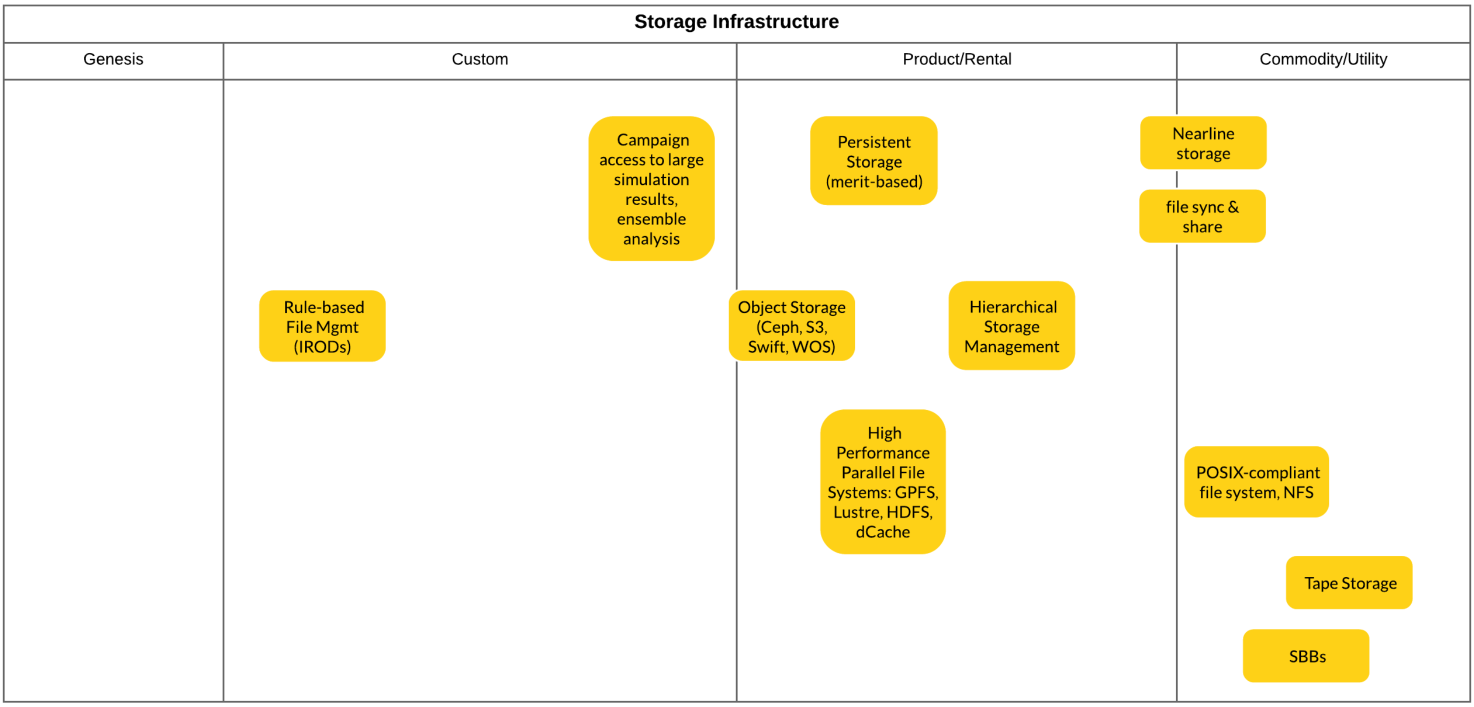
Data movement, to enable ongoing computations and data publication and sharing, is a primary factor when considering cloud vs. on premise storage solutions, and this in turn is driven by the most efficient compute solution rather than any intrinsic characteristics of storage infrastructure. Additional factors are cost, long term access and preservation strategies.

In addition to issues around location, movement and costs, the features required by researchers from related storage services are evolving as the requirements of research data management evolve. For example, every piece of research data should ideally have content-related metadata as well as information specifying who can find the data, who can access that data, how long it should be kept, etc. All of this information needs to be stored, and questions of visibility, access, *expiration date* are intrinsically managed by storage services, so there is a tight relationship between the functionality expected by research data management, and the functions actually provided by underlying storage services.

Exponential increases in research data volume translate into predictable increases in storage investment required over time. This places a new and growing financial burden on research infrastructure facilities, a burden that was not significant in the past.

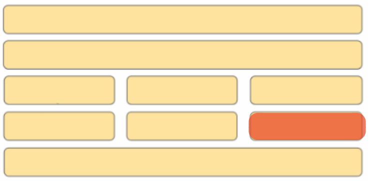
Increasing focus on data preservation adds to this growing volume and these costs, since multiple copies of an entire data set must be stored and/or georeplication software must be used to minimise storage requirements while ensuring integrity of the data. Preserved data must also be regularly checked for integrity and occasionally older file formats need to be converted to new formats to maintain accessibility, all of which drive recurring costs beyond power to make sure preserved data is in fact preserved. However, the exponential pace at which new data is generated means that the volume of preserved older data should still represent just a fraction of the new data volume.

Value Chain of Subcomponents of Storage Infrastructure


Storage Infrastructure
Genesis
Custom
Product/Rental
Commodity/Utility
Campaign access to large simulation results, ensemble analysis
Persistent Storage (merit-based)
Nearline storage
file sync & share
Rule-based File Mgmt (IRODs)
Object Storage (Ceph, S3, Swift, WOS)
Hierarchical Storage Management
High Performance Parallel File Systems: GPFS, Lustre, HDFS, dCache
POSIX-compliant file system, NFS
Tape Storage
SBBs

Network Infrastructure

Network infrastructure links compute and storage infrastructure with the ultimate user. As with storage Infrastructure, the big data explosion places greater demands on the network, since larger data sets are being created in many places and being accessed and processed by others in still more places. Network infrastructure supports direct data transfer activities, remote access and increasingly, data streaming from observational sites to data storage and processing facilities.



Key Trends

Point to point data transfer capacity is growing through advances in optical transport technologies and some expansion of terrestrial and undersea cable connectivity. The pace of performance improvement (roughly 17% per year) is slower than found for compute (roughly 40% per year). For example it took roughly 15 years for transport technologies in general availability to advance from 10 Gbp/s to 100 Gbp/s. The latest transport standard readily supports up to 100 Gbp/s transmissions speeds over a variety of physical media and configurations (although exclusively optical media over longer distances). 400 Gbp/s is regarded as the practical limit of the current standard – higher speeds, such as the terabit Ethernet (TbE) sought by Google and other major players, will require agreement on new standards and development and deployment of new technologies.

The need for higher data rates is driven by the *big data explosion*, specifically by the need to move large data sets from their point of creation, to multiple places where the data can be processed or otherwise used by researchers, and to support secure data storage and preservation. As file sizes increase, at a given transport speed, transfer times increase proportionally, requiring researchers to explicitly budget time for data movement as part of many research activities.

Data movement is a particular challenge across water boundaries – undersea cable is more expensive to install and typically lags terrestrial capacities because undersea components must operate reliably over long periods, without any opportunity for replacement of failed components. Unlike international links between neighbouring countries (e.g. between Canada and the US), cross sectional *bandwidth* between two countries separated by an ocean are much more limited, and transfer times may even have to be scheduled explicitly.

The cost structure for network infrastructure may need to evolve. In most jurisdictions, there is a *national research and education network* (NREN) funded by government to support the data communications needs of local researchers and educators, including interconnection with neighbouring jurisdictions and/or submarine networks, and for most NRENs, users can transmit data for free. When NRENs and their networks were initially established, high speed data networks and connections were not generally available, so the NREN was addressing a market failure, and recovering the cost of the network through incremental transfer fees would have been prohibitively expensive for users. In most developed nations today, high speed data connectivity is broadly available from commercial providers at reasonable prices. In some cases NRENs are using government funding to purchase network services from those commercial providers in order to meet their commitments (e.g. enabling researchers to transfer data from university to university). Arguably domestic transfers of research data could entail cost-recovery fees, although this has not been the norm.

International connectivity, particularly across water boundaries, continues to be challenging, and NRENs play a critical role in enabling this connectivity, through facilitation, management as well as direct investment. Recovering these costs from users may still be prohibitive even today.

Network connections with, and the movement of data to and from, cloud facilities are another area of change and evolution. Initial proposals from commercial cloud service providers usually included incremental charges for the movement of data into (*ingress*) and out of (*egress*) cloud-based data storage. More recently cloud service providers have dropped these charges – at least for academic research users, but the future cost structure for this component of cloud-based processing is hard to predict.

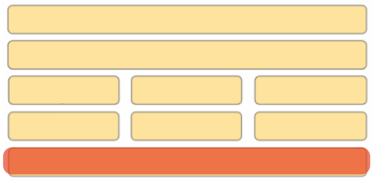
As with both storage and compute infrastructure, basic data transport capabilities are accessed through service layers that provide a variety of functions. High speed file transfer protocols are available from a number of commercial providers as well as *open-source* providers such as Globus, most based on *GridFTP* originally developed for the high energy physics research community. Endpoint configurations such as a *Science DMZ* complement these link-optimising protocols by minimising firewall-related security delays for authorised transfers.

As data volumes, data sources and the value of timely data access increase, ingest of real-time or near-real-time data from multiple, potentially thousands of, sources represents a growing challenge for network infrastructure, straining bandwidth, routing capacity and security mechanisms. For these kinds of many-to-one situations, the traditional paradigm of the *network connects to the data centre* will be joined by an *edge computing* paradigm, where appropriate processing functions can be performed in the network itself rather than in the data centre. For example, data reduction is a common challenge in eResearch projects – large amounts of data are collected from experiments or from the real world, which then must be processed to extract the features of interest. If such features can be defined, the process of extracting those features might be performed closer to the source or at an intermediate processing point, resulting in smaller amounts of data being sent on to the next step, and potentially making the process more efficient and/or less expensive.

Value Chain of Subcomponents of Network Infrastructure

Network Infrastructure			
Genesis	Custom	Product/Rental	Commodity/Utility
		Science DMZ Increased performance File Transfer Data network -- advanced	
	400 Gbps	100 Gbps 40 Gbps	10Gbps

Authentication & Identity Management



All of the previous components need authentication and identity management services to make sure only authorised users are accessing and using the higher level capabilities. This component also incorporates the technology, physical arrangement and business processes required to secure eResearch facilities.

Key Trends

Unlike in the commercial world, legitimate users of eResearch facilities and services can come from many organisations domestically and internationally. An overarching objective for this component is reliably identifying legitimate users and authorising their access to valuable resources without either creating onerous authentication processes (that would slow down valuable research) or requiring (almost impossibly) a centralised identity management system – while at the same time giving researchers the ability *single sign on* – use a single authentication process to establish access any systems on which that researcher was authorised.

The primary trends in support of this objective are:

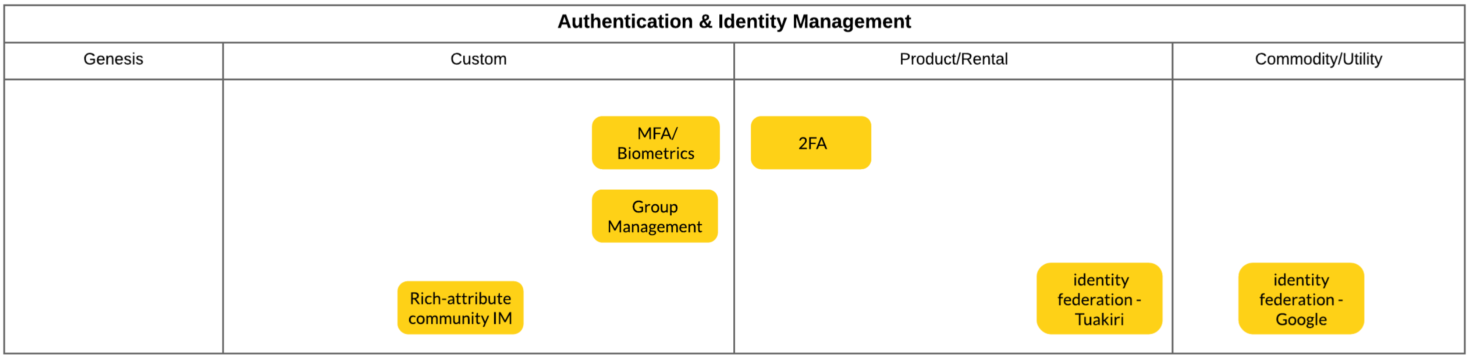
- Expanding federation and inter-federation efforts.
- Integration of more robust mechanisms for authenticating users, particularly through the use of multiple proofs of identity, or factors (*multi-factor authentication*).
- Proliferation of schemes for managing authorisation – what a given researcher is authorised to *do* across the universe of eResearch facilities.

Federated identity management separates the process of authentication (performed by an *identity provider*) from the process of authorisation (performed by a *service provider*). Identity providers release only enough identity information to enable the service provider to authorise (or deny) access to the requested resource. The primary identity management federations are inCommon in the US, and eduGain in Europe and elsewhere, and they in turn federate with each other and with other IM federations. Nevertheless considerable coordination is required for full interoperation of these systems, and for service providers to fully rely on federated identities.

The greater the value of the resource being accessed, the greater the certainty of identity proof required before allowing access. Simple authentication schemes might be acceptable for low value resources (e.g. access to WiFi, downloading lists of readily available information), while more complex authentication might be desired to run 10,000-core analysis jobs on the latest advanced research computing system.   While federated identity management is simplifying the user experience through potential single sign on, multi-factor authentication confounds this simplicity with a growing range of additional candidate factors, such as SMS-transmitted verification codes and biometrics.

Every researcher participates in a variety of research activities that may confer a wide range of access rights – rights flowing from their role in the university, relationship to each of their students, participation in collaborative international projects, and receipt of various funding and resource awards. Various tools are being developed, such as Grouper, to assist with management of these rights, and their potential intersections/overlaps, while also maintaining privacy.

Value Chain of Subcomponents of Authentication & Identity Management



Authentication & Identity Management

Genesis	Custom	Product/Rental	Commodity/Utility
	MFA/ Biometrics	2FA	
	Group Management		
	Rich-attribute community IM	identity federation - Tuakiri	identity federation - Google

Appendix A:

Overview of New Zealand Research

The Government's vision for 2025 is of:

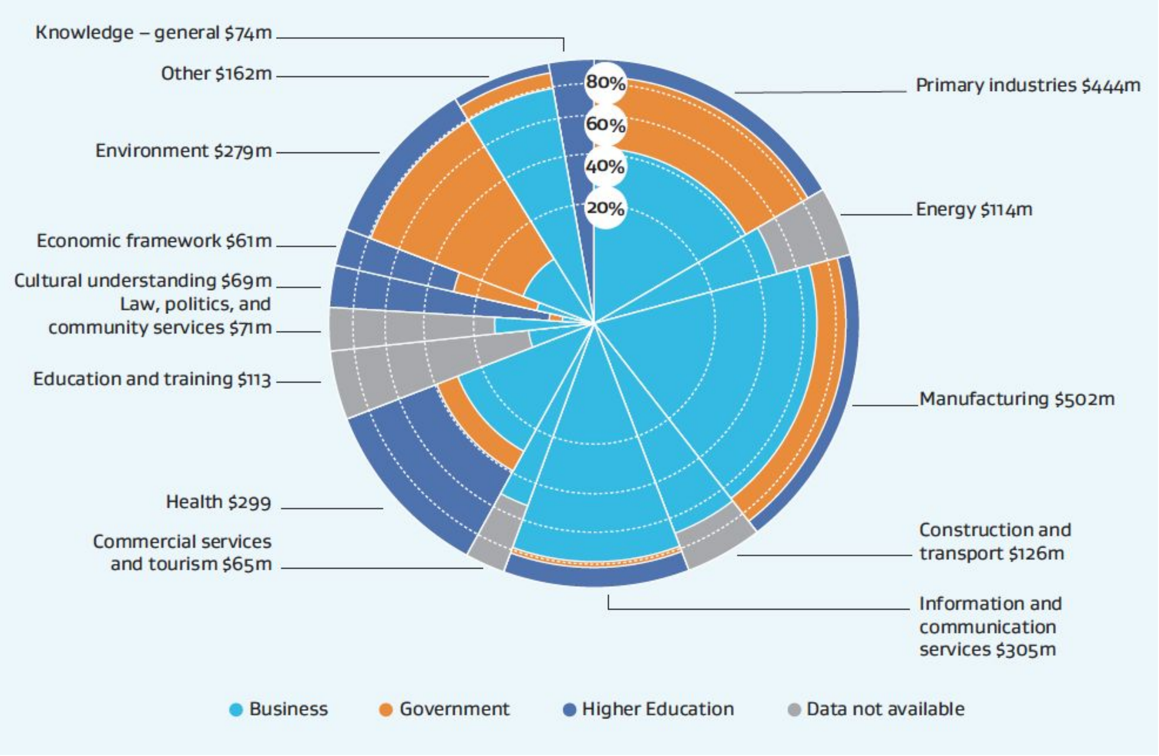
> *A highly dynamic science system that enriches New Zealand, making a more visible, measurable contribution to our productivity and wellbeing through excellent science.*

National Statement of Science Investment (NSSI) 2015-2025

The NSSI stated that health and basic ICT research were a priority. The NSSI identified that future growth in primary sector R&D should be driven by industry, with government support.

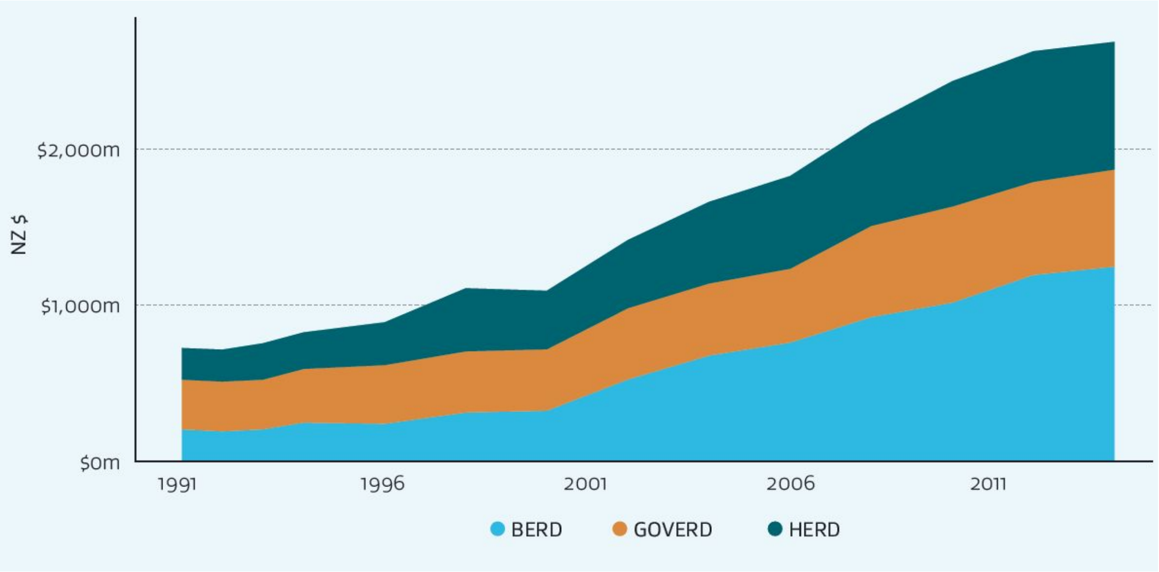
The focus of the NSSI is on two pillars: excellence and impact (economic, environment, social, health).

Expenditure on R&D by purpose of research and sector of expenditure (2014)

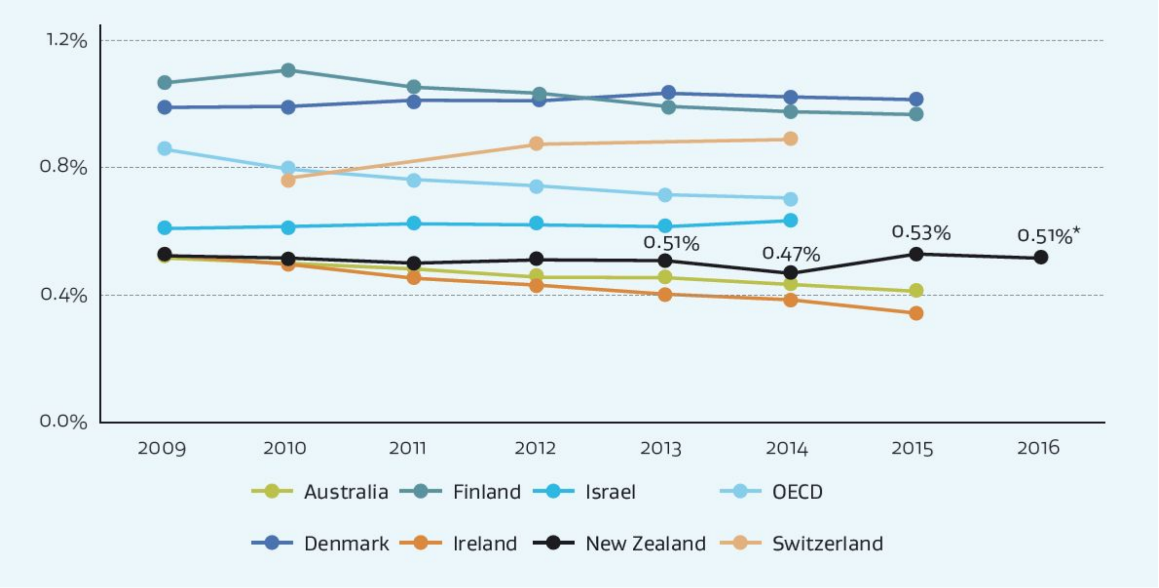


R&D Funding and Expenditure

R&D growth by sector of performance, nominal NZ $



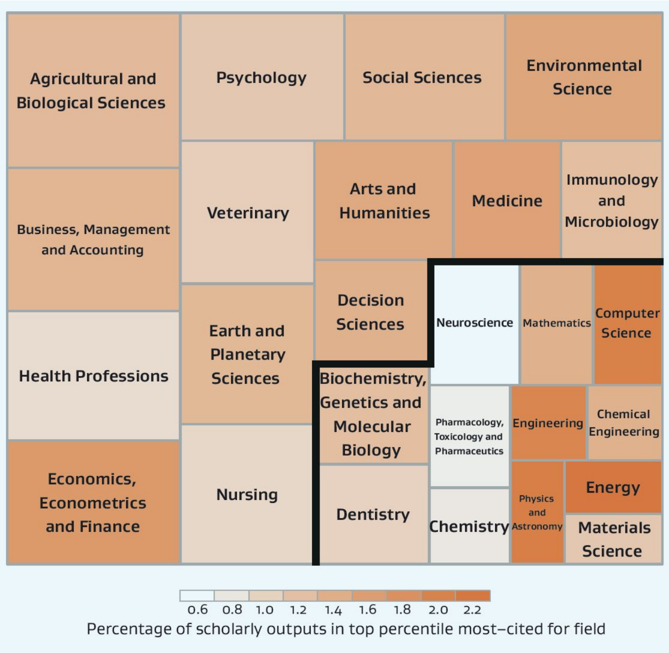
Public funding of R&D as a proportion of GDP



New Zealand's public funding of R&D is lower than the OECD average, is on a par with Ireland, Australia and Israel, but is significantly less than Denmark and Finland (which are among the highest in the world). Government's investment (in both real and nominal terms) has risen significantly over this period. The economy has also grown over this time, which has resulted in the percentage figure staying relatively constant. Business expenditure on R&D is still relatively low when compared with other Small Advanced Economies and the OECD.

New Zealand's scientific specialisations[1]

New Zealand's revealed comparative advantage in research volume (size of box) and publications in top 1% most-cited for the field (shading), 2010-13.
New Zealand publishes a higher share of its research outputs than the world average in fields above and to the left of the thick line.



New Zealand publishes over half its output in just five fields (Medicine; Agricultural and Biological Sciences; Social Sciences; Biochemistry, Genetics and Molecular Biology; and Engineering). However, New Zealand's relative specialisms are in a broader range of fields than the simple output data would suggest. The top nine (rather than five) fields constitute just over half of the total 'weighted output'. Agriculture remains but Medicine no longer figures as a relatively large share of New Zealand's output. Other specialisms are revealed including in Business, Management and Accounting; Veterinary; Health Professions; Psychology; and Economics. These are fields in which New Zealand does relatively more research than the OECD average. New Zealand has some really excellent research – publications which appear in the top 1 percent of cited work for that field worldwide. Now, another set of very niche specialisms is revealed – in Engineering; Physics and Astronomy; Computer Science; and Energy research. New Zealand's current focus of research effort on Agriculture, Veterinary, Health Professions, Environmental Science, and Earth and Planetary Sciences probably reflects New Zealand's economy, society and environmental (including geological) challenges.

The niche expertise in areas relating to technology and IT suggest opportunity for these to contribute to economic diversification in these high productivity sectors. Basic ICT research is an area where the NSSI committed to increasing investment over time.

New Zealand research sector workforce

Role	Higher Education	Crown Research Institutes
Researchers	10,700	1,860
Technicians	2,000	558
Students	16,400	0
Total	29,100	2,557
Percent	**92**	**8**

[1] 2016 Science and Innovation System Performance Report: http://www.mbie.govt.nz/info-services/science-innovation/performance/system-performance-report

eScience Infrastructure in New Zealand

Strategic Science Investment Fund

The Strategic Science Investment Fund (SSIF) supports underpinning research programmes and infrastructure of enduring importance to New Zealand. The SSIF is a major lever to deliver on the vision of the National Statement of Science Investment - a better-performing science system that is larger, more agile and more responsive, investing effectively for long-term impact.

The SSIF supports funding for national research infrastructure platforms that provide access to research technology, facilities, infrastructure, Nationally Significant Collections and Databases, and associated support services.

eResearch 2020

eResearch 2020 conducted a qualitative study launched by NeSI, REANNZ and NZGL in 2014. We spoke with researchers across a wide range of disciplines, as well as leaders of ICT and management at our research institutions. In each case we asked them to discuss the potential for change in their work over the coming decade, and to highlight the opportunities and challenges they see before us as a national research system.

The aspiration is for the New Zealand research system to be functioning as a best-in-class small country sector in 2020.

"Inevitably, our society's problems in the future are going to have a data and computation aspect to them."

Professor Peter Hunter

In the coming decade, New Zealand is likely to have one shot at developing digital infrastructure at a national scale in our research sector. We need to begin that development process with cross-sector research communities, as well as with individual institutions, if we aim to design the best outcomes.

eResearch 2020 Recommendations

- Provide training and adjust incentives to address the growing skills gap in our research system when it comes to digital methodologies and research quality.
- Enable and promote sharing of risk and investment into the infrastructure and capability layer in our research system. This includes promoting cohesiveness and knowledge exchange within national research communities, rather than research institutions.
- Take a national eco-system view of digital research capability and how we meet changing expectations in research quality.
- Endeavour to be self-sufficient in those strategic circumstances that require it, and actively invest to ensure the underlying human capability to understand new technologies is never lost to us.

New Zealand eScience Infrastructure (NeSI)

NeSI was established in 2011 to provide a nationally coordinated high-performance computing (HPC) network to enhance the capability and quality of New Zealand's research. NeSI's establishment was in response to an undersupply of supercomputing facilities in New Zealand and was intended to overcome coordination failure in New Zealand's science system. NeSI was established through an investment partnership across the Crown and five research institutions collectively investing NZ$48 million over 4 years to support both infrastructure and team.

NeSI's second phase of investment commenced in 2014 with combined funding NZ$53 million to June 2018. NeSI implemented a number of recommendations flowing from an evaluation conducted in 2013, including, among other things, restructuring into a nationally led unitary team, leading the design and procurement of future infrastructure, growing its reach to support a greater breadth of users, and integrating cloud and data management strategies with existing HPC investment. NeSI's contract has been extended to June 2019 to allow development of an appropriate follow-on business case and funding structure.

Recently, NeSI made a capital investment of roughly NZ$11 million, partnering with the National Institute of Weather and Atmospheric Research Ltd (NIWA) which has made its own capital investment of NZ$8 million, to purchase and commission two new advanced research computers:

- Capability system, Maui, a Cray CX50 supercomputer with 18,560 Skylake cores, 66.8 TB of total memory, ARIES Dragonfly interconnect (estimated performance: 1,425 Tflops)
    - Attached pre- and post-processing system, 1,200 Skylakes cores, 8 Nvidia GPUs, 23 TB of total memory
- Capacity system, Mahuiki, a Cray CS400 cluster with 8,424 Broadwell cores, 30 TB of total memory, Infiniband interconnect (estimated performance: 308 TFlops)
    - Attached pre- and post-processing system, 640 Broadwell cores, 8 Nvidia GPUs, 12 TB of total memory
    - Cloud-like infrastructure delivery supported via OpenStack on a portion of nodes
    - Support for Cloud-bursting
- Shared high performance data storage, an IBM Spectrum Scale and ESS storage solution

Both systems should be ready for general use by mid 2018. NIWA's investment entitles it to use 43% of the Maui system to support NIWA's operational research programs -- the balance will be available to New Zealand researchers through NeSI's access and allocation processes.

Research and Education Advanced Network New Zealand

Research and Education Advanced Network New Zealand (REANNZ) operates New Zealand's national research and education network (NREN). NeSI relies on the network provided by REANNZ, and uses Tuakiri, a federated identity management service provided by REANNZ, for user authentication and authorisation. NeSI has had a close relationship with REANNZ since its foundation in 2006.

Genomics Aotearoa

Genomics Aotearoa (GA) is a new collaborative platform for genomics and bioinformatics in NZ established in August 2017. NeSI is working with GA to determine how NeSI can best support GA as it works to meet researcher needs in this field. In the past NeSI had coordinated services with a predecessor to GA called New Zealand Genomics Limited (NZGL), whose funding was not renewed.

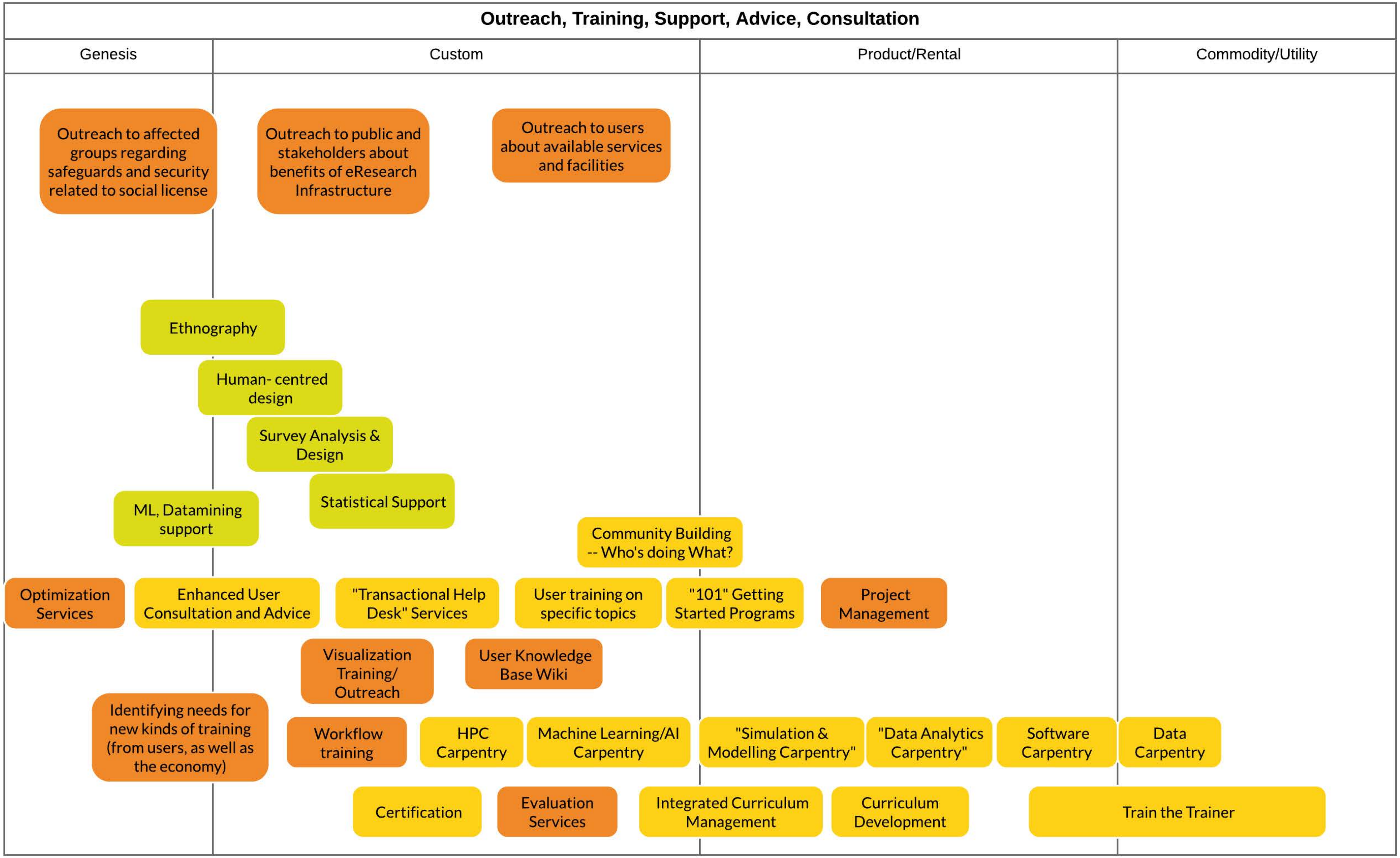
Appendix B:

Value Chain Component Maps and Descriptions of Subcomponents

This section is a compilation of all value chain component maps introduced in the main report, scaled to fit the page. Following each map is a table with descriptions for each subcomponent.

For all following value chain maps, each component has been *coded* in three colours to reflect discussions during and in parallel with the workshops, as well the ongoing refinement of the maps themselves:

- Yellow components were included in the preliminary value chains presented in the workshop.
- Orange components were identified by attendees as required.
- Green components represent important services and activities brought up in other presentations at the event.



Outreach, Training, Support, Advice, Consultation

Genesis

Custom

Product/Rental

Commodity/Utility

Outreach to affected groups regarding safeguards and security related to social license

Outreach to public and stakeholders about benefits of eResearch Infrastructure

Outreach to users about available services and facilities

Ethnography

Human- centred design

Survey Analysis & Design

Statistical Support

ML, Datamining support

Community Building -- Who's doing What?

Optimization Services

Enhanced User Consultation and Advice

"Transactional Help Desk" Services

User training on specific topics

"101" Getting Started Programs

Project Management

Visualization Training/ Outreach

User Knowledge Base Wiki

Identifying needs for new kinds of training (from users, as well as the economy)

Workflow training

HPC Carpentry

Machine Learning/AI Carpentry

"Simulation & Modelling Carpentry"

"Data Analytics Carpentry"

Software Carpentry

Data Carpentry

Certification

Evaluation Services

Integrated Curriculum Management

Curriculum Development

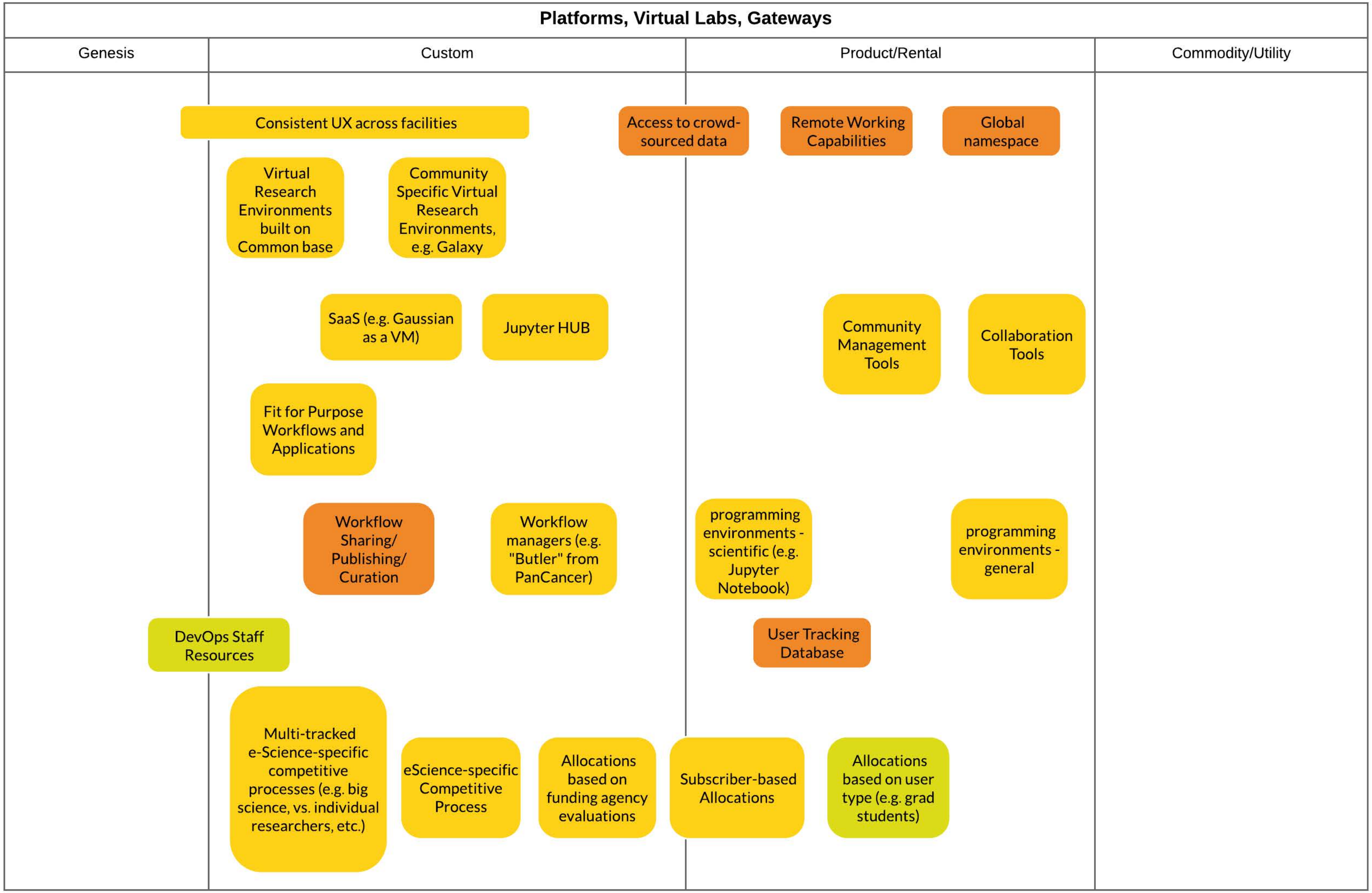
Train the Trainer

Outreach, Training, Support, Advice, Consultation

Component Title (Function, Service, Activity)	Description
"101" Getting Started Programs	Basic training for users to get started and become productive with capabilities and services of facility
"Data Analytics Carpentry"	Training in basic principles and practical techniques for analysis of large amounts of data to identify patterns, correlations and anomalies in that data
"Simulation & Modelling Carpentry"	Training in basic principles and practical techniques of computational simulation and modelling of systems
"Transactional Help Desk" Services	On demand assistance for users encountering difficulties in the use of the facility
Certification	Test to certify that individuals possess minimum knowledge required to be "proficient" in a specific activity.
Community Building -- Who's doing What?	Knowledge bases, wikis, databases, etc. allowing researchers to learn how other researchers are using advanced research computing to advance their research.
Curriculum Development	Develop training programs to ensure trainees acquire the knowledge required to be "proficient" in a specific activity.
Data Carpentry	Training in basic principles and practical techniques of data analysis. "Data Carpentry develops and teaches workshops on the fundamental data skills needed to conduct research. Our mission is to provide researchers high-quality, domain-specific training covering the full lifecycle of data-driven research. "
Enhanced User Consultation and Advice	Consultation and advice for users, potentially on a longer term basis (e.g. over weeks or months), assisting users in achieving their research objectives using the facility
Ethnography	Consultation and advice from ethnographers, trained at the intersection of sociology, anthropology, and computer and data sciences, to help ensure stakeholder alignment, governance, and other organisational and communication-related frameworks that will contribute to successful eResearch projects.
Evaluation Services	Processes and resources for evaluating the skill set and skill levels of individuals, to assess success of training activities or to diagnose the need for new skills training.
Fit for Purpose Workflows and Applications	Knowledge bases, wikis, databases, etc. allowing users to search for and find software workflows and applications to perform certain research-related functions
HPC Carpentry	Training in basic principles and practical techniques for use of high performance (capability) computing, particularly employing large numbers of compute processors, integrated through tightly coupled interconnection networks.
Human- centred design	Consultation and advice from design specialists to improve usability (UI - web, applications, mobile) and accessibility (ensuring both adherence to standards and accommodation of people with different abilities) to, among other things, improve adoption and reduce time and complexity of use.
Identifying needs for new kinds of training (from users, as well as the economy)	Processes and resources to identify gaps in training; for example, capturing this data during helpdesk sessions, or conducting surveys with industry to identify workforce training gaps.
Integrated Curriculum Management	Design content of training programs so that one program takes advantage of knowledge acquired in any "pre-requisite" programs, and to so that content is not unnecessarily duplicated.
Machine Learning/AI Carpentry	Training in basic principles and practical techniques of machine learning and artificial intelligence.
ML, Data mining support	Specialised assistance for research employing machine-learning and/or data-mining techniques and tools.

Optimisation Services	Services whereby appropriately skilled staff can review codes, workflows and research environments in order to improve performance or effectiveness.
Outreach to affected groups regarding safeguards and security related to social license	Communications activities directed at groups who "own" or have an interest in certain kinds of data, information and knowledge to increase awareness of the processes and safeguards that are in place to ensure ethical and proper use of that data and conduct of related research.
Outreach to public and stakeholders about benefits of eResearch Infrastructure	Communications activities directed at the public and specific stakeholders to increase awareness of how eResearch Infrastructure creates benefits to society and to selected stakeholders.
Outreach to users about available services and facilities	Communications activities directed at users and potential users to increase awareness of available services and facilities, particularly how each service or facility might be valuable to a given group of users.
Project Management	Project management is the practice of initiating, planning, executing, controlling, and closing the work of a team to achieve specific goals and meet specific success criteria at the specified time. In the context of eScience, project management comprises additional activities: integrating the work of a research team with the work of eScience practitioners, as well as eScience facilities, to meet the research project objectives; applying specialised knowledge of eScience techniques, resources and facilities to improve the outcomes of the research project, including offering advice to project leaders regarding alternative approaches to the project that may help improve the project's outcomes.
Software Carpentry	Training in basic principles and practical techniques for development of efficient research software that can be tested to perform as intended under a wide variety of conditions, and can be maintained and updated over time. "Software Carpentry has been teaching researchers the computing skills they need to get more done in less time and with less pain."
Statistical Support	Assistance in the use of statistical techniques in designing and interpreting analyses and simulations.
Survey Analysis & Design	Assistance in the design of effective survey tools and instruments and in the valid analysis of survey data collected.
Train the Trainer	Execute programs to teach individuals wishing to conduct training in a subject how to train others in that subject
User Knowledge Base Wiki	Technology for sharing user knowledge in a collaborative way, augmenting traditional helpdesk or support functions.
User training on specific topics	More advanced training on specific capabilities and services
Visualisation Training/ Outreach	Training in basic principles and practical techniques in the use of visualisation as a tool used during research analyses. Communication to users and potential users to increase awareness of visualisation that can help researchers be more productive in their research activities.
Workflow training	Training in basic principles and practical techniques of the use of workflows to perform certain kinds of research analyses.

Platforms, Virtual Labs, Gateways



Genesis

Custom

Product/Rental

Commodity/Utility

Consistent UX across facilities

Access to crowd-sourced data

Remote Working Capabilities

Global namespace

Virtual Research Environments built on Common base

Community Specific Virtual Research Environments, e.g. Galaxy

SaaS (e.g. Gaussian as a VM)

Jupyter HUB

Community Management Tools

Collaboration Tools

Fit for Purpose Workflows and Applications

Workflow Sharing/ Publishing/ Curation

Workflow managers (e.g. "Butler" from PanCancer)

programming environments - scientific (e.g. Jupyter Notebook)

programming environments - general

DevOps Staff Resources

User Tracking Database

Multi-tracked e-Science-specific competitive processes (e.g. big science, vs. individual researchers, etc.)

eScience-specific Competitive Process

Allocations based on funding agency evaluations

Subscriber-based Allocations

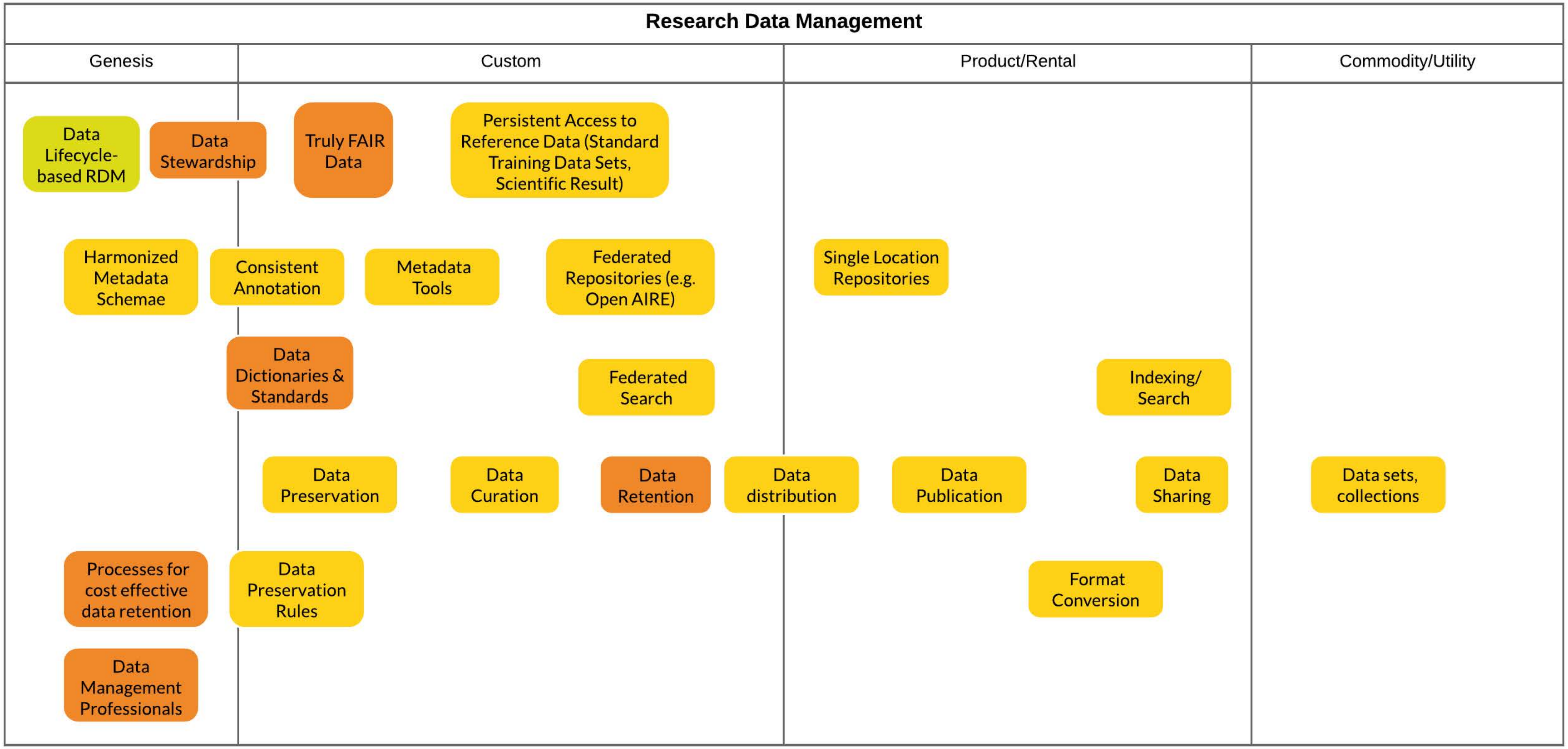
Allocations based on user type (e.g. grad students)

Platforms, Virtual Labs, Gateways

Component Title (Function, Service, Activity)	Description
Allocations based on funding agency evaluations	Facility resources (e.g. storage capacity, compute capacity) are allocated to research users who have requested access in a way that reflects the quality of the research to be performed using those resources, as evaluated by funding agencies that have separately reviewed the research in order to determine funding levels.
Allocations based on user type (e.g. grad students)	Facility resources (e.g. storage capacity, compute capacity) are allocated to some research users according to user type -- e.g. grad students may automatically receive set allocations of storage and/or compute capacity, newly appointed faculty may receive larger allocations, etc. User types and "default" allocation amounts would be set to reflect relevant science priorities.
Collaboration Tools	Common software tools to enable computer-based collaboration on projects
Community Management Tools	Common software tools to support interactions of a group of like-minded individuals, or individuals with a common purpose, defining and identifying community members and managing the ways those members can interact with others
Community Specific Virtual Research Environments, e.g. Galaxy	Software system, integrating a number of components, including data sets and repositories, application software and workflows, collaborative communications tools, and visualisation or display tools into a computer-based "environment" or workspace that allows researcher to perform research computing activities more easily and efficiently. Specific VREs have been created by different research disciplines to meet pressing productivity needs, without any attempts to find common, discipline-agnostic, solutions to common functional requirements. For example, computational job submission procedures could reflect the specific ARC systems available to the original developers, rather than attempting to interface to more general job scheduling systems.
Consistent UX across facilities	Configuration and use of the same or similar sets of middleware across multiple ARC facilities, allowing users to more rapidly become productive in the use of any of those facilities.
Container-based virtualisation, such as Docker	Containers formalise and productise the concept of virtual machine images
DevOps Staff Resources	Technology professionals skilled and experience in a software engineering culture and practice that aims at unifying software development (Dev) and software operation (Ops). DevOps skills are valuable when there are rapid iterations between development and operation, which is typical in a research environment.
Docker Swarm/ Kubernetes/ Mesos	Software tools for managing a collection of container-based configurations in order to interoperate on a multi-node ARC hardware system.
eScience-specific Competitive Process	Facility resources (e.g. storage capacity, compute capacity) are allocated to research users who have requested access based on the quality of the research to be performed using those resources, as evaluated by the eScience facility itself in a competitive, peer-reviewed process.

Global namespace	Technologies by which electronic files (data, software) can be identified and accessed without reference to specific facilities or systems.
Jupyter HUB	System that supports the creation of multiple single-user Jupyter Notebook instances, in order to enable multi-user access to Jupyter Notebook.
Multi-tracked eScience-specific competitive processes (e.g. big science, vs. individual researchers, etc.)	Facility resources (e.g. storage capacity, compute capacity) are allocated to research users who have requested access based on the quality of the research to be performed using those resources, as evaluated by the eScience facility itself in a competitive, peer-reviewed process. Research users apply for such resources in different categories of access, enabling fairer comparison of dissimilar types of user proposals (e.g. separating major collaborative science projects from single researcher based projects)
OpenStack	Widely accepted software system that allows multi-node servers to be partitioned into smaller "virtual machines" (VMs) and made available to users so that they are not aware of other users of the larger system and can use their virtual machine as if it were a separate system dedicated to their use. Virtualisation imposes some performance penalties on the user, since the virtualisation software processes requests from the user and passes them on to the operating system of the larger host system. Users benefit from virtualisation by having greater (apparent) control over their virtual machine, for example allowing them to install software that might conflict with software installed on the host system.
Puppet	Software tool for automating configuration of ARC hardware (compute nodes, CPU and node interconnect, etc.) in order to execute a specific ARC job.
Remote Working Capabilities	Process and resources to allow researchers to conduct digitally-enabled research activities from anyplace in the world and not requiring their physical presence to access related research data, tools or facilities.
SaaS (e.g. Gaussian as a VM)	"Software as a service" making common software applications easier to use, e.g. allowing a user to submit materials for processing and then receive back the desired results.
Singularity/ Shifter	Software tools for managing a tightly connected (e.g. through Infiniband) collection of container-based configurations in order to interoperate on a multi-node ARC hardware system.
SLURM, Torque/ Maui/ MOAB, PBS, UNIVA Grid Engine	Software tools for specifying ARC "jobs" (analysis software, hardware configuration, data to be staged for processing, user accounts and resource allocations to be consumed) and submitting them to one or more ARC facilities for execution consistent with the facility's scheduling and prioritisation policies.
Subscriber-based Allocations	Organisations that have contributed financially to a facility ("subscribers") are entitled to set amounts of facility resources (e.g. storage capacity, compute capacity) and allocate some or all of those entitlements to users associated with the organisation, according to their own priorities and procedures.
UNICORE, gLite, HTCondor	Software tools for finding resources available to execute single- and multi-node "jobs" on multiple multi-node ARC systems. Distributed execution differs from scheduling because the execution software deterministically decides which jobs should run when, while a scheduler manages competition for resources from job requesters that have varying levels of "priority" access to those resources.

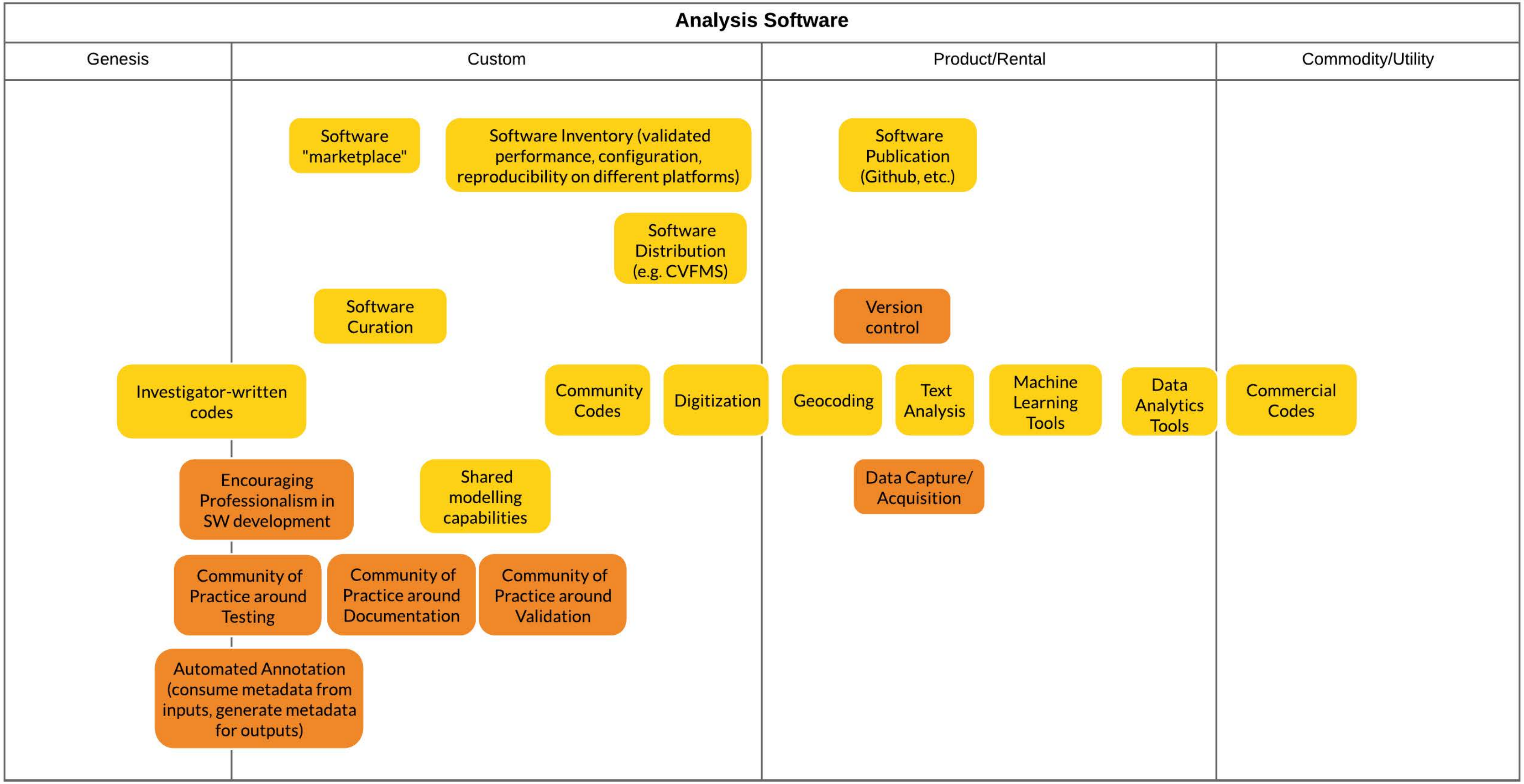
User Tracking Database	Repository of information regarding facility users and related data, such as allocation and usage data, relationships with other users and/or membership in groups.
Virtual Research Environments built on Common base	Software system, integrating a number of components, including data sets and repositories, application software and workflows, collaborative communications tools, and visualisation or display tools into a computer-based "environment" or workspace that allows researcher to perform research computing activities more easily and efficiently. VREs built on a common base take advantage of common requirements (e.g. data management, workflow definitions, computational job submission, visualisation) by using the same tools to meet these requirements, regardless of the research discipline, or the unique VREs that may have been created in the past in different disciplines.
VM Image Libraries	Collection of previously configured virtual machine "images", combining analysis software and required software libraries or interfaces, any of which can be quickly installed, and reliably operated, in a virtual machine environment such as OpenStack.
Workflow managers (e.g. "Butler" from PanCancer)	Software tools to define multi-step research computing tasks (data retrieval, processing, storage, conditional step execution, error or failure handling), initiate such tasks and track their operational progress, including failures.
Workflow Sharing/ Publishing/ Curation	Processes and resources for sharing, publishing and curating analysis workflows, allowing users to apply validated multi-step analysis processes to new data, or to reproduce analysis results produced from data collected and shared by others.


Research Data Management
Genesis
Custom
Product/Rental
Commodity/Utility
Data Lifecycle-based RDM
Data Stewardship
Truly FAIR Data
Persistent Access to Reference Data (Standard Training Data Sets, Scientific Result)
Harmonized Metadata Schemae
Consistent Annotation
Metadata Tools
Federated Repositories (e.g. Open AIRE)
Single Location Repositories
Data Dictionaries & Standards
Federated Search
Indexing/ Search
Data Preservation
Data Curation
Data Retention
Data distribution
Data Publication
Data Sharing
Data sets, collections
Processes for cost effective data retention
Data Preservation Rules
Format Conversion
Data Management Professionals

Research Data Management (evolving to Research Artifact Management)

Component Title (Function, Service, Activity)	Description
Access to crowd- sourced data	Technology supporting electronic data capture from multiple individuals
Consistent Annotation	Ensuring consistent annotation of large data collections from different sources by multiple annotaters, combining the effective use of metadata tools and well-defined work procedures
Data Curation	Human-based process for supervising and directing the annotation, preservation, sharing and preservation of research data. Data curators ensure that data under their purview is findable, accessible, interoperable and re-usable (FAIR).
Data Dictionaries & Standards	Dictionary: Repository of information about data such as meaning, relationships to other data, origin, usage, and format (ref: IBM Dictionary of Computing). Standards: Documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data.
Data distribution	Allowing an authorised user (typically an "owner" or "custodian") to distribute a data set or data record to multiple authenticated users ("sharers") on a permanent or time-limited basis. This function does not allow the sharing users to modify the original data set or record. Sharers receive a copy of the shared data automatically, without having to request a copy.
Data Lifecycle- based RDM	Processes and resources that encourage the effective management of research data over its complete lifecycle -- from initial generation/creation to eventual permanent archiving or destruction.
Data Management Professionals	Individuals trained the practice of data management who can assist others with data stewardship
Data Preservation	Processes and resources that ensure that valuable data sets are preserved so that they can be found and accessed over long periods of time.
Data Preservation Rules	Rules for the periodic "refresh" of archived file formats, ensuring that preserved research data are accessible, readable and usable by researchers using relatively current software. A "library" analogy would be scanning older films into digital video formats.
Data Publication	Allowing an authorised user (typically an "owner" or "custodian") to make a data set or data record "public" (copyable by) on a permanent or time-limited basis. Sharers may be required to be authenticated or may be anonymous. This function does not allow the sharing users to modify the original data set or record. Sharers access public data by requesting to copy the data.
Data Retention	Processes and resources that ensure that data sets are stored, without corruption, and accessible to authorised users for a defined retention period.
Data sets, collections	Collections of data stored in multiple files, potentially in multiple formats, as well as in databases.
Data Sharing	Allowing an authorised user (typically an "owner" or "custodian") to make a data set or data record accessible to (copyable by) other authenticated users ("sharers") on a permanent or time-limited basis. This function does not allow the sharing users to modify the original data set or record. Sharers access shared data by requesting to copy the data.
Data Stewardship	Factors encouraging creators and owners of research data to care for that data in a way that creates value for others.
Federated Repositories (e.g. Open AIRE)	Repositories in multiple locations that provide common features and functions, and operate according to the same principles, such that users can access the data from any location without needing to access that location individually.

Federated Search	Software tools for indexing diverse sets of data stored in multiple repositories, using metadata associated with each data set and/or database entry, and for making those indices searchable by authorised users to find data meeting the user's search criteria.
Format Conversion	As part of a data preservation capability, archived research data are periodically converted from older file formats to newer formats according to Data Preservation Rules.
Harmonised Metadata Schemae	Coordination of metadata standards so that data that should be annotated using multiple standards can be consistently annotated.
Indexing/Search	Productised software (e.g. Elastic Search) for indexing diverse sets of data in a single repository, using metadata associated with each data set and/or database entry, and for making those indices searchable by authorised users to find data meeting the user's search criteria.
Metadata Tools	Software tools to assist researchers when annotating data according to one or more metadata standards, e.g. allowing simultaneous annotation of multiple data sets with common metadata information, or ensuring selection of metadata field values from approved drop down lists.
Persistent Access to Reference Data (Standard Training Data Sets, Scientific Result)	Selected data sets are stored on a long term basis by a facility and shared with a broad user group.
Processes for cost effective data retention	Processes that set retention periods for data that balance the societal cost of retention (including both storage as well as ongoing preservation and curation activities) against the potential future value of that data
Single Location Repositories	Data collections stored in a single facility, enhanced with features including preservation, annotation of data sets or records with metadata in one or more specific formats, user authentication, data access control according to general user rights or sharing/release rules that vary for each record or data set.
Truly FAIR Data	Processes and resources that ensure that data sets are easily Findable, Accessible, Interoperable and Reusable.



Analysis Software

Genesis

Custom

Product/Rental

Commodity/Utility

Software "marketplace"

Software Inventory (validated performance, configuration, reproducibility on different platforms)

Software Publication (Github, etc.)

Software Distribution (e.g. CVFMS)

Software Curation

Version control

Investigator-written codes

Community Codes

Digitization

Geocoding

Text Analysis

Machine Learning Tools

Data Analytics Tools

Commercial Codes

Encouraging Professionalism in SW development

Shared modelling capabilities

Data Capture/ Acquisition

Community of Practice around Testing

Community of Practice around Documentation

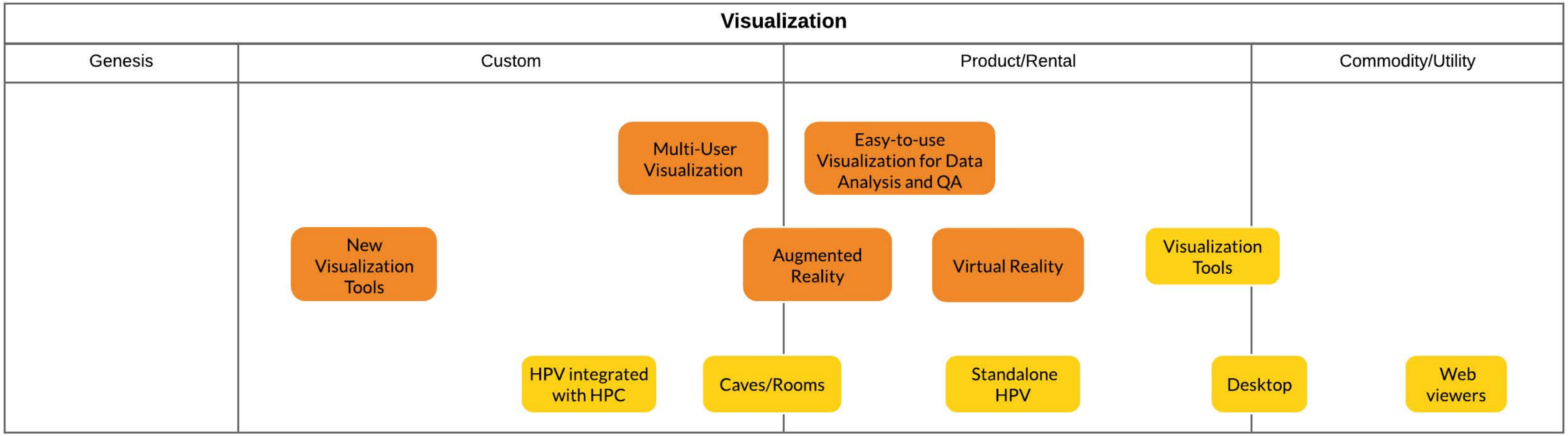
Community of Practice around Validation

Automated Annotation (consume metadata from inputs, generate metadata for outputs)

Analysis and Modelling Software

Component Title (Function, Service, Activity)	Description
Automated Annotation (consume metadata from inputs, generate metadata for outputs)	Optional feature of analysis software that annotates output files with metadata based on the metadata associated with input files and based on the parameters of the analysis run that creates the outputs.
Commercial Codes	Software systems sold/licensed for commercial use.
Community Codes	Software systems developed by research communities, usually on an open-source basis, typically offering limited formal user-support services, but also typically supported by the research community itself.
Community of Practice around Documentation	Software developers working together as an informal community to promote and support the use of improved software documentation practices.
Community of Practice around Testing	Software developers working together as an informal community to promote and support the use of improved software testing practices.
Community of Practice around Validation	Software developers working together as an informal community to promote and support the use of improved software validation practices.
Data Analytics Tools	Software tools for analysis of large data sets, e.g. available through AWS
Data Capture/ Acquisition	Technology for reliably receiving, formatting and storing high volume data generated by experimental or observational systems.
Digitisation	Machine-learning and/or crowd-sourced (Mechanical Turk/Captcha) capture of digital information from still- or moving-images or audio.
Encouraging Professionalism in SW development	Factors encouraging greater use of professional practices by developers of research software
Geocoding	Machine-learning and/or crowd-sourced (Mechanical Turk/Captcha) annotation of information with geolocation data.
Investigator-written codes	Software systems developed by one or a few researchers, usually on a proprietary basis, intending to accomplish a research function not well addressed by community or commercial codes or with better performance than those codes.
Machine Learning Tools	Productised software tools using the principles of machine learning to identify patterns or features in general data.
Programming environments - general	Interactive software system that allows software developers to develop software more easily and efficiently.
Programming environments - scientific (e.g. Jupyter Notebook)	Interactive software system that allows researcher to develop research computing software more easily and efficiently.
Shared modelling capabilities	Common frameworks for high resolution simulation modeling that allow for interoperability of the many software components required for effective performance and accuracy. Weather and climate modeling is the principal domain where shared modeling capabilities are being developed.
Software "marketplace"	Computer-based listing of software tools, applications and related components available to perform a variety of research-related tasks, allowing users to access and use appropriate tools, possibly for a fee, to achieve their individual research objectives.

Software Curation	Human-based process for supervising and directing what software is needed for different purposes and in different electronic locations. For example, software curators would ensure that any older software code or components required to properly access and manipulate published data sets is also available for use, as well as managing different versions of more current software tools so that users of older versions are properly supported or encouraged to migrate to more current versions.
Software Distribution (e.g. CVFMS)	Computer-based process for installing specific versions of software code or components on multiple research computing systems so that users of those systems can access the correct versions of the software in support of their research efforts. Software distribution can eliminate the need for a researcher to access a software marketplace, or to use a software inventory, and instead ensure that the required software is correctly installed on the systems needed.
Software Inventory (validated performance, configuration, reproducibility on different platforms)	Computer-based listing of software tools, applications and related components available to perform a variety of research-related tasks, providing users with instructions on how to access and use each tool, possibly for a fee. Inventory provides more detail on versions, electronic locations and component-dependencies than a "marketplace," while not necessarily providing direct access to or authorisation to use each tool.
Software Publication (Github, etc.)	Database of software tools, applications and related components that can be combined and compiled to perform a variety of research-related tasks. Publication focusses on access to versions of code and/or compiled components with defined provenance and performance, possibly linked to specific reference or published data sets and research results, enabling independent reproduction and validation of those results.
Text Analysis	Software tools for analysis of text, most commonly frequency analysis (e.g. Word maps), word and phrase correlations (e.g. distance)
Version control	Software development processes (human processes) related to the creation, documentation and publication of successive versions of software components, particularly when multiple developers are involved in this process. Version control software assists the version control process, enable activities such as reversion, when an older version of a software component must be returned to active use because a newer version is not ready for use.



Visualization

Genesis

Custom

Product/Rental

Commodity/Utility

Multi-User Visualization

Easy-to-use Visualization for Data Analysis and QA

New Visualization Tools

Augmented Reality

Virtual Reality

Visualization Tools

HPV integrated with HPC

Caves/Rooms

Standalone HPV

Desktop

Web viewers

Visualisation

Component Title (Function, Service, Activity)	Description
Augmented Reality	Visualisation technique that overlays a user's current environment with additional visual information, typically through the use of goggles and other special purpose "wearable" technology. Examples include "head up displays" of combat information in a military environment, or equipment condition and maintenance information in an industrial environment.
Caves/Rooms	Special-purpose rooms, equipped with video projectors or arrays of video displays, that enable immersive (3D) visualisation of large data sets. Game-style controls allow exploration of the data.
Desktop	Computer-game style visualisation and exploration of large data sets on a single desktop computer, often supported by dedicated data processing on separate servers.
Easy-to-use Visualisation for Data Analysis and QA	Easy to use visualisation tools that can be readily set up and used by researchers wanting to use visualisation as a quick tool for data analysis or quality assurance.
HPV integrated with HPC	Co-location of high performance visualisation capabilities with high performance compute systems, enabling visualisation of large HPC output data sets without requiring transfer of that data to other systems. This capability is particularly valuable when visualisation is used to visually -- and quickly -- validate results of modelling or analyses.
Multi-User Visualisation	Visualisation tools that allow multiple users to experience the same visualisation at the same time.
New Visualisation Tools	Early stage higher-performance visualisation tools, most likely at the experimental stage.
Standalone HPV	Dedicated high-performance servers for visualisation of large data sets stored in scratch or "project" storage systems of an ARC facility.
Virtual Reality	Visualisation technique that immerses the user in a realistic 3D environment, typically through the use of goggles and other special purpose "wearable" technology.
Visualisation Tools	Productised software for processing of large data sets to create still or moving image visualisations of selected data.
Web viewers	Visualisation and exploration of large data sets using web-based interfaces and navigation controls, usually supported by server-side data processing to generate and update images.

Compute Infrastructure

Genesis	Custom	Product/Rental	Commodity/Utility
Adaptable Configurations Massive InMemory Storage (a la COSMOS)	HPC applications service Graph/ Data Analytics High Availability Compute (for NIWA: Data mirroring, plus fast failover of compute to DR facility) VM Image Libraries Singularity/ Shifter Puppet Increased performance Large-scale Capability (>1000 core) Increased performance Mixed Node Types	SLURM, Torque/ Maui/ MOAB, PBS, UNIVA Grid Engine Container-based virtualization, such as Docker Docker Swarm/ Kubernetes/ Mesos OpenStack Medium-scale capability (~1000 cores) 16-core capability Large-scale Capacity (>1000 core) Medium-scale capacity (~1000 cores) 16-core capability VM Server interconnect (Infiniband, Ethernet, OPA, etc.) Large Memory Compute Nodes Uniform node-types SSDs	16-core capacity RAM

Compute Infrastructure

Component Title (Function, Service, Activity)	Description
16-core capability	On premises hardware configuration of 16 tightly coupled cores
16-core capability VM	Cloud-based configuration of 16 tightly coupled cores
16-core capacity	Cloud-based configuration of 16 loosely coupled cores
16-core capacity	On premises hardware configuration of 16 loosely coupled cores
Adaptable Configurations	Software or middleware allowing virtual or logical nodes to be configured in different ways, allowing ARC jobs requiring different hardware configurations to be executed effectively, while at the same time not "wasting" nodes that do not have the right configuration to run any pending ARC job.
Graph/ Data Analytics	Server configurations designed specifically to execute graph or data analytics ARC jobs more efficiently.
High Availability Compute (for NIWA: Data mirroring, plus fast failover of compute to DR facility)	Hardware systems configured so that submitted ARC jobs are guaranteed to be executed with a high probability, even if certain hardware or environmental failures occur.
Increased performance	
Increased performance	
Large-scale Capability (>1000 core)	On premises hardware configuration of more than 1000 tightly coupled cores
Large-scale Capacity (>1000 core)	Cloud-based configuration of more than 1000 loosely coupled cores
Large-scale Capacity (>1000 core)	On premises hardware configuration of more than 1000 loosely coupled cores
Large Memory Compute Nodes	Hardware servers configured with large amounts of integrated NVM. Typical "large memory" configurations are on the order of 0.5-3.0 TB
Massive InMemory Storage (a la COSMOS)	Hardware servers configured with access to massive amounts of shared memory. The COSMOS system at Cambridge has roughly 1,850 CPU nodes accessing a common store of over 14 PB of globally accessible shared RAM. (The global accessibility of this memory is the unique feature -- most memory is accessible only by the processor or node to which it is attached.)

Medium-scale capability (~1000 cores)	On premises hardware configuration of roughly 1000 tightly coupled cores
Medium-scale capacity (~1000 cores)	Cloud-based configuration of roughly 1000 loosely coupled cores
Medium-scale capacity (~1000 cores)	On premises hardware configuration of roughly 1000 loosely coupled cores
Mixed Node Types	Multi-node server where different nodes have different configurations, allowing ARC jobs requiring different hardware configurations to be executed effectively.
RAM	Random access memory accessible to the processor or node to which it is attached.
Server interconnect (Infiniband, Ethernet, OPA, etc.)	Hardware allowing high-speed direct communications between processors and nodes of a multi-core ARC system
SSDs	Solid State Disks -- commodity storage hardware using solid state memory rather than spinning disk technology, but packaged to appear to other systems either as a traditional SBB with higher performance, or as RAM connected to related processors or nodes by the NVMe interface
Uniform node-types	Multi-node server where each node has the same configuration, allowing execution of multi-node, high core-count ARC jobs.

Storage Infrastructure			
Genesis	Custom	Product/Rental	Commodity/Utility
	Campaign access to large simulation results, ensemble analysis Rule-based File Mgmt (IRODs)	Persistent Storage (merit-based) Object Storage (Ceph, S3, Swift, WOS) Hierarchical Storage Management High Performance Parallel File Systems: GPFS, Lustre, HDFS, dCache	Nearline storage file sync & share POSIX-compliant file system, NFS Tape Storage SBBs

Storage Infrastructure

Component Title (Function, Service, Activity)	Description
Campaign access to large simulation results, ensemble analysis	Longer term access to the very large output data files (uncompressed)resulting from large-scale simulations and/or suites of data files resulting from simulations executed against a suite of input parameters ("parameter sweeps").
file sync & share	Convenient user interface (e.g. drag and drop) for advanced storage functionality
Hierarchical Storage Management	Software that manages the placement of data files on different storage systems in order to minimise cost per unit of storage while maintaining acceptable latency for file access. For example, data expected to remain at rest ("write once, read never") could be stored on tape without affecting user satisfaction. Infrequently accessed data files could be stored on low-cost, low-performance spinning disk. Frequently access data (such as reference data sets) might be stored on SSD.
High Performance Parallel File Systems: GPFS, Lustre, HDFS, dCache	File storage systems optimised to work with multiple, typically tightly coupled, compute servers. Features include simultaneous, yet synchronised, reads and writes at different points within a single large data file, as well as varying levels of failure tolerance/data restoration, and possibly HSM and more-rudimentary file-locking features.
Nearline storage	Storage system providing high latency access to infrequently accessed data files. Technology could be tape or low-performance, low-cost spinning disk systems
Object Storage (Ceph, S3, Swift, WOS)	Integrated storage management system that combines features of Hierarchical Storage Management, as well as backup, georeplication and unified namespace features.
Persistent Storage (merit-based)	Longer term (project, campaign or semi-archived) storage of data files.
POSIX-compliant file system, NFS	Traditional File Storage systems found in enterprise systems running UNIX or Microsoft operating systems.
Rule-based File Mgmt (IRODs)	Flexible mechanism for providing the features of an Object Storage System using a rules-based engine such as iRODS.
SBBs	Storage Building Blocks -- commodity spinning magnetic disk storage hardware providing various capacities and performance (latency, throughput)
Tape Storage	Storage of data on magnetic tapes, packaged in cartridges or cassettes that are themselves stored in robotic tape libraries, enabling multi-petabyte data storage capacities at lower cost (per byte) than disk storage, but with higher time to access (latency) due to the sequential nature of tape reads and writes.

Network Infrastructure			
Genesis	Custom	Product/Rental	Commodity/Utility
	400 Gbps	Science DMZ Increased performance File Transfer Data network -- advanced 100 Gbps 40 Gbps	10Gbps

Network Infrastructure

Component Title (Function, Service, Activity)	Description
100 Gbps	Raw network bandwidth generally available across a network of ARC production nodes (individual node-to-node bandwidth - not network cross section bandwidth).
10Gbps	Raw network bandwidth generally available across a network of ARC production nodes (individual node-to-node bandwidth - not network cross section bandwidth).
40 Gbps	Raw network bandwidth generally available across a network of ARC production nodes (individual node-to-node bandwidth - not network cross section bandwidth).
400 Gbps	Raw network bandwidth generally available across a network of ARC production nodes (individual node-to-node bandwidth - not network cross section bandwidth).
Data network -- advanced	Integration of transport functions with higher level functions such as file transfer, authorised access to secure DMZ facilities, etc.
File Transfer	Services to transfer large data files to/from/between ARC facilities, typically using FTP or more preferably GridFTP or similar variations, typically terminating at specially configured data transfer nodes at ARC production facilities dedicated to managing the transfer process and then storing the data set at the intended location (e.g. persistent project storage to which the transferring user has authorised access and sufficient resource rights)
Science DMZ	Hardware and network systems configured to increase file transfer speeds by reducing throughput delays associated with network firewalls (due to packet examination ("sniffing") activities). Since cybersecurity could be compromised through such configurations, access to a Science DMZ should be limited to authorised users.

Authentication & Identity Management			
Genesis	Custom	Product/Rental	Commodity/Utility
	MFA/ Biometrics Group Management Rich-attribute community IM	2FA identity federation - Tuakiri	identity federation - Google

Authentication & Identity Management

Component Title (Function, Service, Activity)	Description
2FA	2 factor authentication -- requiring 2 pieces of private information (password and one other factor) to authenticate a user
Group Management	Within a common identity management scheme, users may form into voluntary groups of users, which can be specified and managed as a type of identity and then used to manage access to resources within the community. For example, climate scientists may form a group within the Tuakiri access federation and then grant access to climate science data to group members, with varying read/write privileges set by the group.
identity federation - Google	Confirming identity using Google sign in services. (I actually don't think this is federated -- i.e. Google does not accept, e.g, Facebook credentials, for authentication)
identity federation - Tuakiri	Software service from REANNZ integrating identity management from multiple institutions (primarily research institutions/higher education) -- allowing users with accounts at any of these institutions to be identified against the credentials provided to those institutions (username and password) and then authenticated for access to services provided by a wide variety of providers. (Authentication does not automatically grant the user rights and/or resource allocations with those service providers; that is a separate process.)
MFA/ Biometrics	Multi-factor authentication -- requiring 2 or more pieces of private information (password and one other factor) to authenticate a user. Could include biometric information such as a a fingerprint or facial scan.
Rich-attribute community IM	Identity federations can assign rights to users based on attributes provided by their "home" identity provider (e.g. this user is a doctoral student and is entitled to a minimum quantum of compute resources, while this user is a PI on this grant and is entitled to a large amount of resources). Members of the community (e.g the institutions) need to trust one another to provide accurate attributes for users, and there must be agreement on the schema for these attributes (even setting "ranks" for academic personnel can be contentious)

Appendix C: List of eScience Futures Workshop Attendees

The authors acknowledge and thank the workshop attendees for participating and contributing in the eScience futures workshop. Their input has helped shape up a common foundation for future community discussions.

Name	Affiliation
Max Wilkinson	eResearch 2020
Dr Murray Poulter	NeSI, NIWA
Stephen Whiteside	NeSI, the University of Auckland
Dr Michael Uddstrom	NeSI, NIWA
Igor Portugal	Catalyst IT
Jana Makar	NeSI
Laura Casimiro	NeSI
Steve Cotter	CSST
Dr Fabiana Kubke	NeSI, the University of Auckland
Professor Cris Print	Genomics Aotearoa, the University of Auckland, ESR
Dave Fellinger	iRODS
Sarah Nisbet	eRSA
Professor Barbara Chapman	NeSI, Stony Brook University
Dr Kim Handley	Genomics Aotearoa, the University of Auckland
Jonah Duckles	Software and Data Carpentries
Professor Andrew Rohl	NeSI, Curtin University
Professor Margaret Hyland	MBIE
Marcus Gustafsson	NeSI, the University of Auckland
Sue Bridger	Microsoft
Associate Professor Jo Lane	the University of Waikato
Jim Donovan	REANNZ
Rick Christie	NeSI
Bill Ritchie	Callaghan Innovation

www.nesi.org.nz