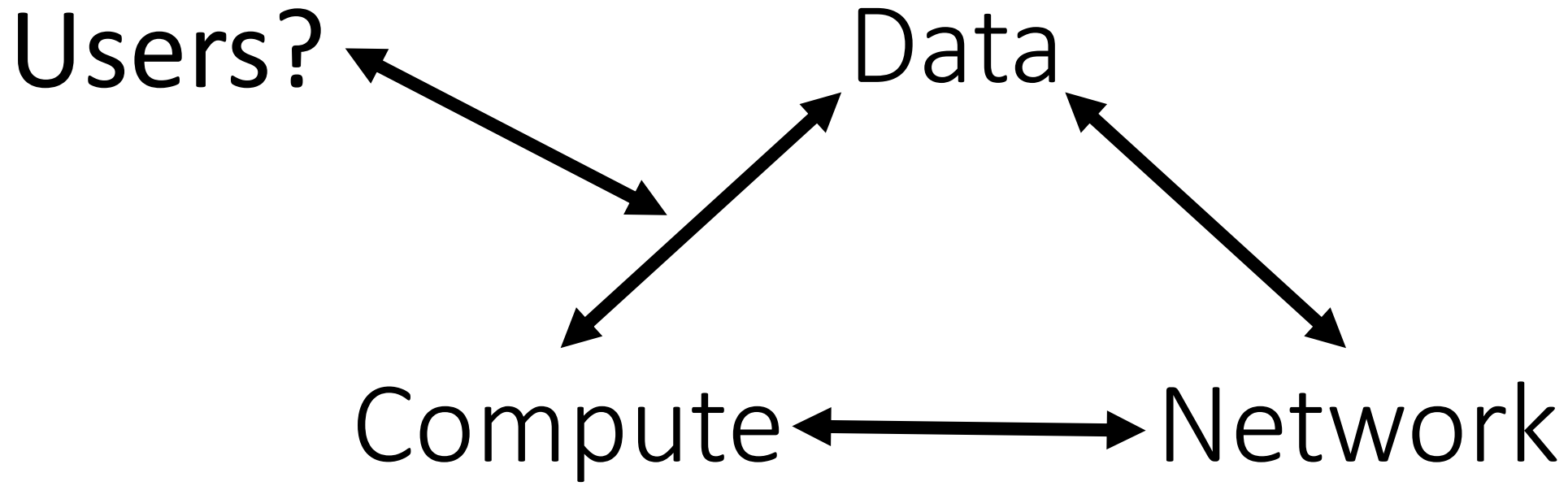# Accelerating Understanding Summit 2016

Moving from compute-centric to data-centric and network-centric – the implications for HPC

Dr Martin Hilgeman, HPC Consultant EMEA - Dell
Dr Thomas Connor, Senior Lecturer - Cardiff University
Dr Herbert Cornelius, Technical Director Advanced Computing EMEA, Intel

Users? ↔ Data ↔ Compute ↔ Network

Implications for HPC in Biology

Dr Thomas R Connor

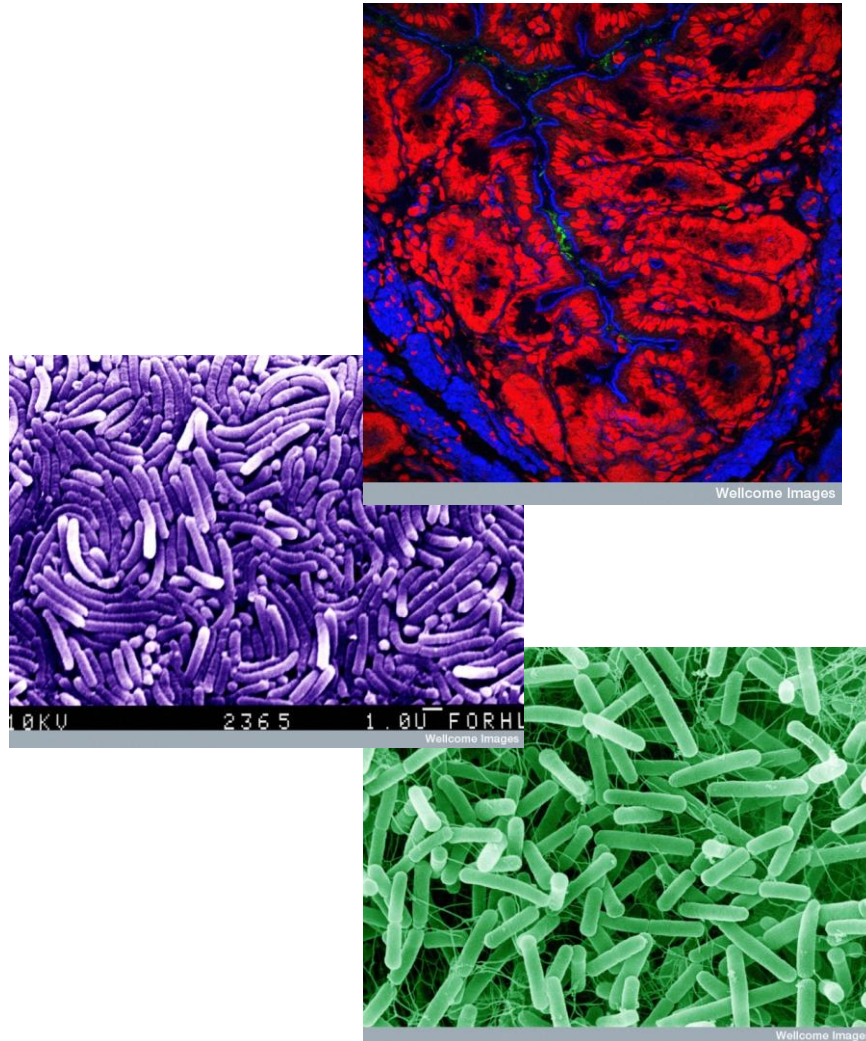School of Biosciences, Cardiff University

connortr@cardiff.ac.uk ; @tomrconnor

The Bigger Picture

- 200 Million people have GI disease at any point in time

- In a day, they will produce ~60,000,000 litres of diarrhoea

- That is equivalent to all the water passing over Victoria Falls in one minute

Why it matters

- 2 Billion cases of disease every year worldwide
- ~5% of all deaths in low and middle income countries are due to diarrhoeal diseases
- Mostly kills children
- GI pathogens are just one of many that collectively pose a serious public health problem
- Working in places where even basic infrastructure isn't available is a key problem
- Makes it very difficult to move the compute to where the data is
- **Key implication is: if we get it right, we can save lives**

What is Genomics

- I am a bioinformatician, and I spend my time working on genomics data from bacterial pathogens

- Have heard the phrase a lot here, but what is Genomics?

- Genomics is a term given to a group of technologies

- These technologies allow us to explore the genome sequence of an organism
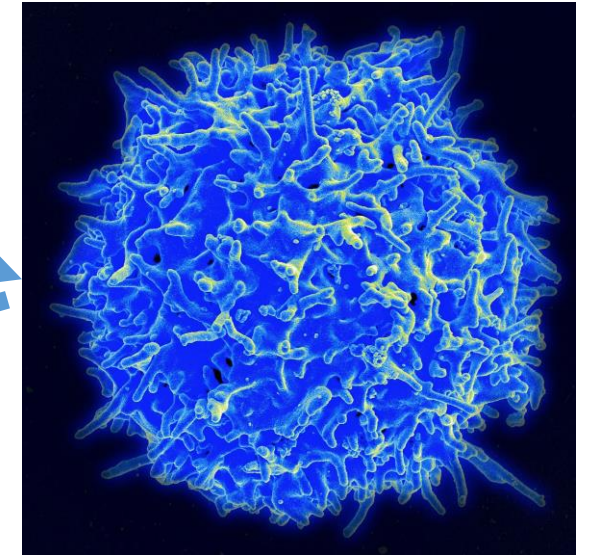

Wellcome Images

What does that mean

DNA encodes the blueprint for virtually every cell of every organism on the planet
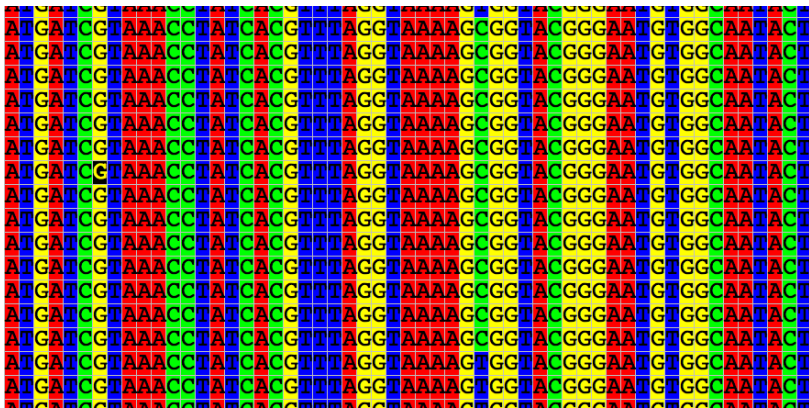
That blueprint defines the features of the cell in which it is found

Genomics enables us to read this blueprint
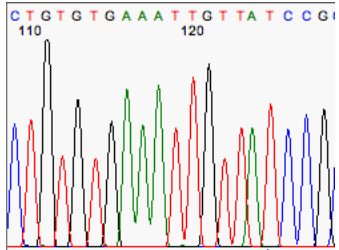
By means of an introduction…..

@IL0_0000:1:1:1:1#0/1

gattatccttcgctcaatctggggcaggcggtgatggtctattgctatcaattagcaacattaatacaacaaccggcgaaaagtgatgcaacggcagacc

+

BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

96 reads/run

Illumina: 500,000,000+ reads/run

454: 1,000,000 reads/run

Nanopore

Pac Bio

Ion Torrent

Solexa / Illumina

Roche 454

ABI "Sanger" sequencers

1995   1998   2000   2003   2008   2010   2012
                                                2011

First Bacterial Genome Sequence

MLST

Draft of the Human Genome

Human Genome finished

Evolution of Typhi (19 genomes)

Evolution of MRSA ST239 (63 genomes)

Evolution of PMEN1 (240 genomes)

1,000 human genomes published

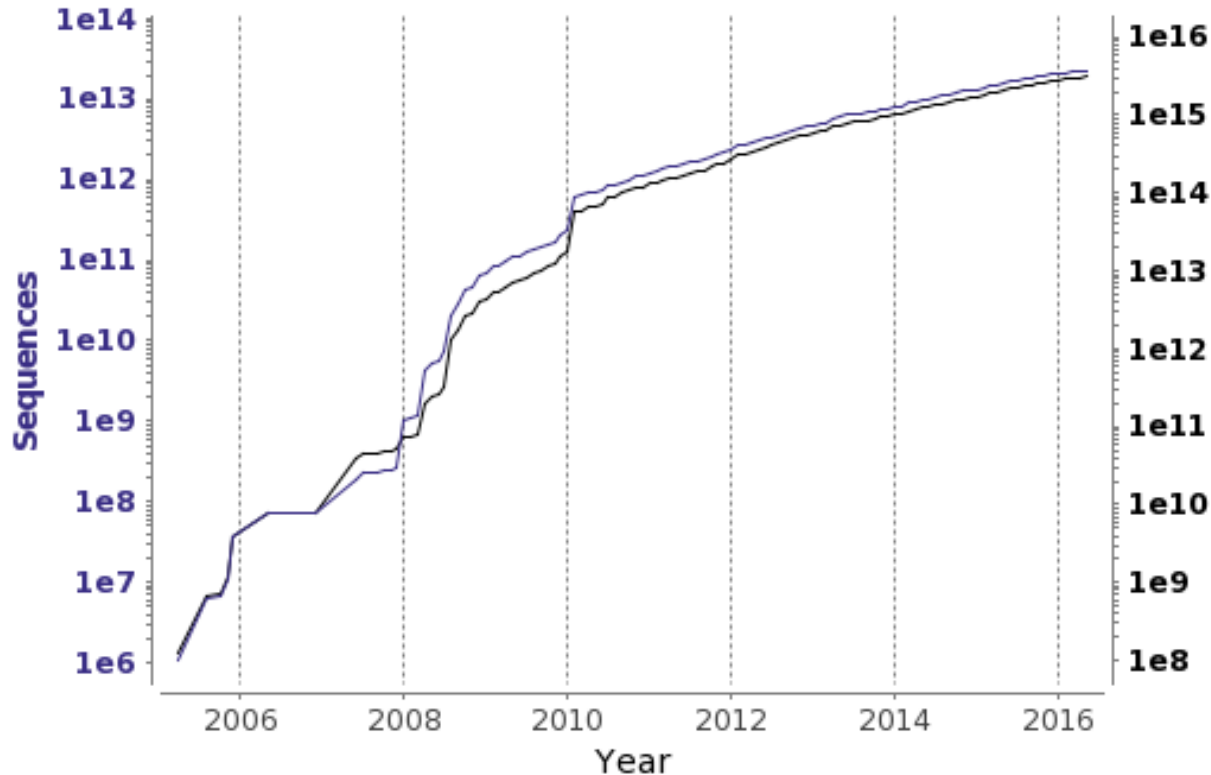# Bottom Line: Sequencing is getting cheaper and easier

2003

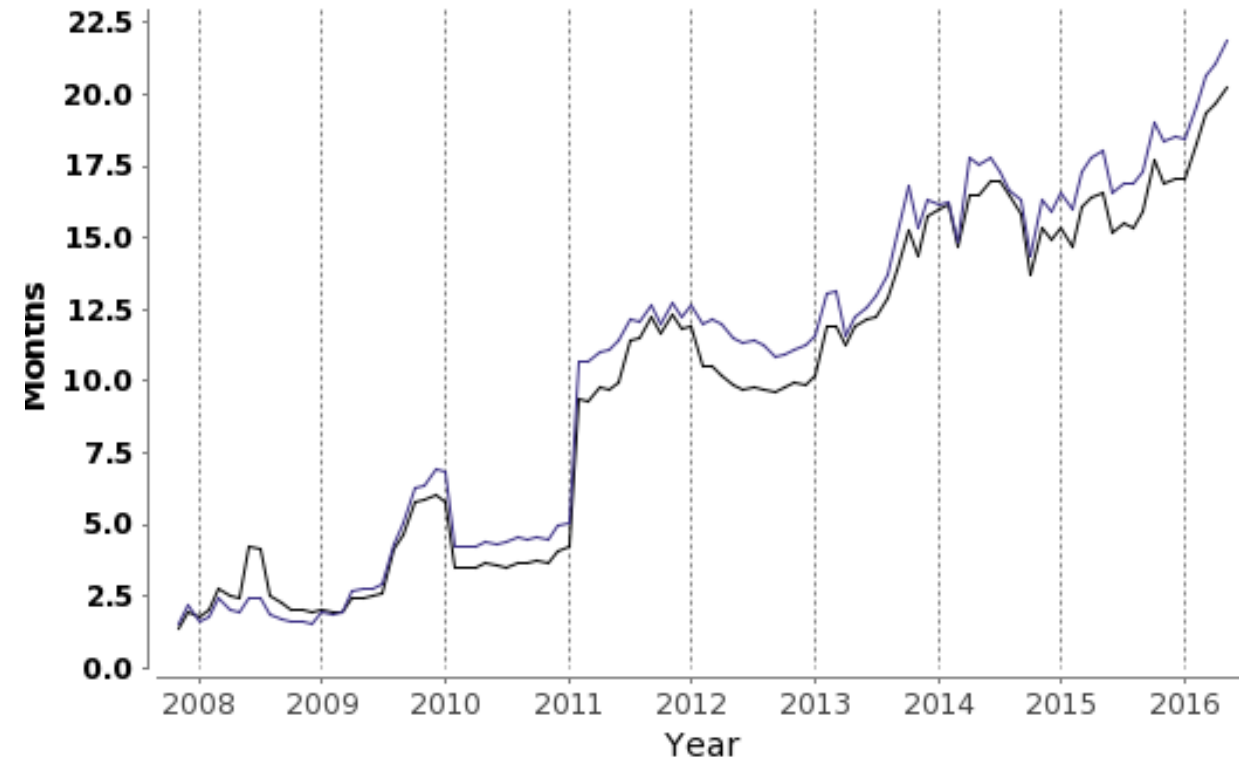2016

Global sequencing data is growing fast

Reads growth
23-May-2016

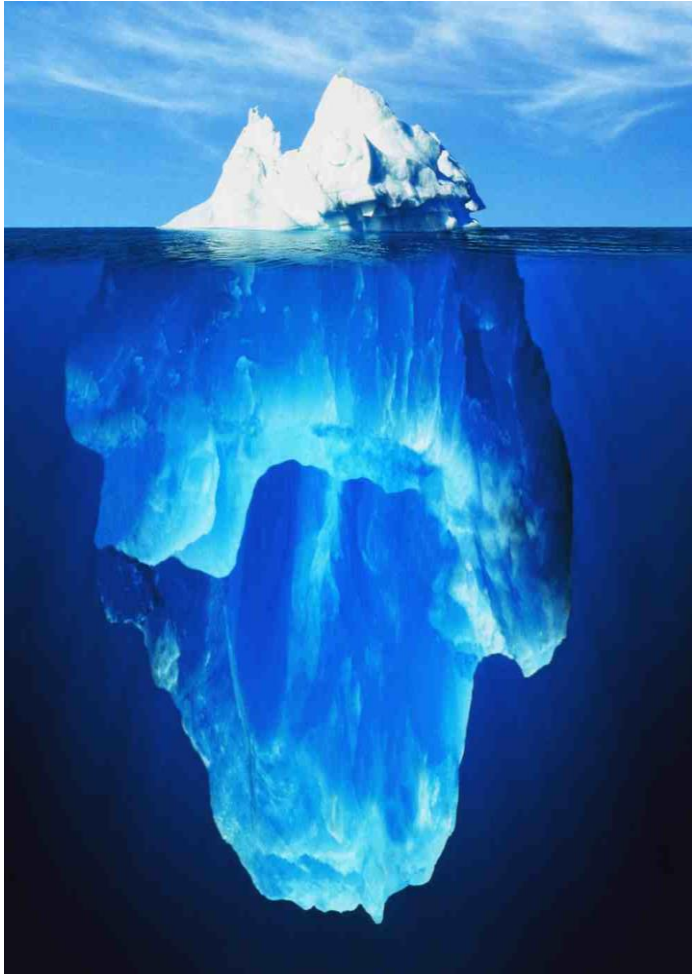**Reads doubling time**
23-May-2016

Sequences (22.6 trillions) — Bases (3,063.4 trillions)

Sequences (21.8 months) — Bases (20.3 months)

Challenges of Biological Big Data

There are many biological analysis platforms available now that make producing large, rich complex datasets relatively cheap and easy

However, the major costs and difficulties do not lie with the generation of data, they lie with how we share, store and analyse the data we generate
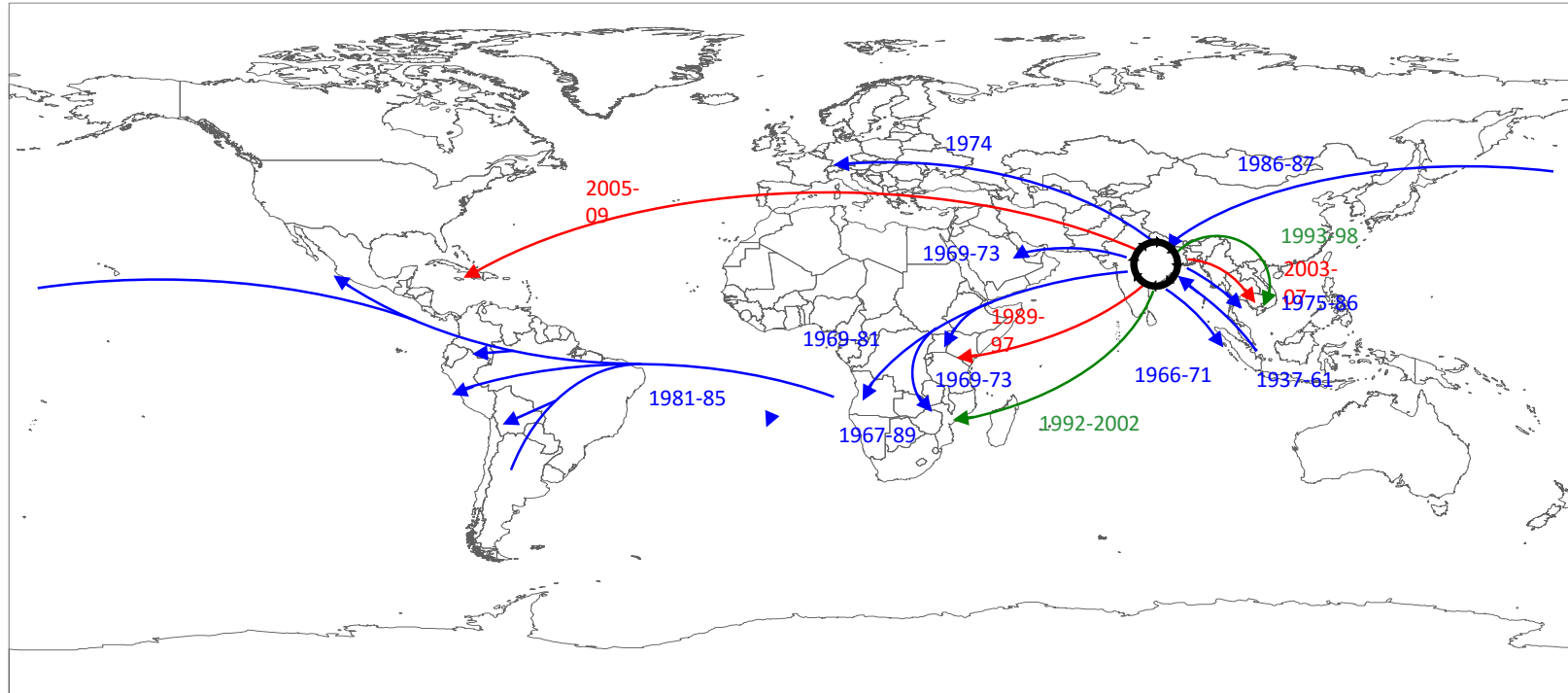
Informatics expertise

Network capacity

User accessibility of software/hardware

Appropriate compute capacity

Software development

Storage availability

The scale of the problem

*Mutreja, Kim, Thomson, Connor et al, Nature, 2011*

At the other end of the scale

```
top - 06:55:06 up 1 day, 20:57,  1 user,  load average: 80.85, 81.31, 81.46
Tasks: 916 total,  82 running, 834 sleeping,   0 stopped,   0 zombie
%Cpu(s): 42.3 us, 42.0 sy,  0.0 ni, 15.6 id,  0.0 wa,  0.1 hi,  0.0 si,  0.0 st
KiB Mem:  29230940+total, 18189089+used, 11041852+free,   102388 buffers
KiB Swap:        0 total,        0 used,        0 free. 92621299+cached Mem
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
13447	ubuntu	20	0	0.637t	0.597t	1020	R	99.6	21.9	10282:37	velvetg
25158	ubuntu	20	0	3127832	2.748g	11372	R	99.6	0.1	540:02.18	Ray
25142	ubuntu	20	0	3127740	2.748g	11356	R	99.3	0.1	539:55.40	Ray
25144	ubuntu	20	0	3127776	2.743g	11372	R	99.3	0.1	540:03.22	Ray
25145	ubuntu	20	0	3127684	2.741g	11364	R	99.3	0.1	540:15.43	Ray
25146	ubuntu	20	0	3127844	2.745g	11376	R	99.3	0.1	540:01.13	Ray
25149	ubuntu	20	0	3127676	2.751g	11384	R	99.3	0.1	539:59.63	Ray
25150	ubuntu	20	0	3128148	2.753g	11360	R	99.3	0.1	539:59.58	Ray
25152	ubuntu	20	0	3124052	2.740g	11364	R	99.3	0.1	539:49.83	Ray
25154	ubuntu	20	0	3127800	2.745g	11360	R	99.3	0.1	540:06.91	Ray
25155	ubuntu	20	0	3127976	2.742g	11380	R	99.3	0.1	540:08.55	Ray
25156	ubuntu	20	0	3127740	2.741g	11372	R	99.3	0.1	540:01.02	Ray
25159	ubuntu	20	0	3124120	2.740g	11380	R	99.3	0.1	540:02.65	Ray
25162	ubuntu	20	0	3127824	2.751g	11380	R	99.3	0.1	539:54.52	Ray
25163	ubuntu	20	0	3127672	2.754g	11364	R	99.3	0.1	539:58.78	Ray
25164	ubuntu	20	0	3127684	2.751g	11392	R	99.3	0.1	540:19.34	Ray
25166	ubuntu	20	0	3127708	2.744g	11388	R	99.3	0.1	539:57.87	Ray
25167	ubuntu	20	0	3127996	2.747g	11372	R	99.3	0.1	540:05.34	Ray
25168	ubuntu	20	0	3127704	2.746g	11396	R	99.3	0.1	540:07.64	Ray
25169	ubuntu	20	0	3127824	2.750g	11364	R	99.3	0.1	540:01.92	Ray
25170	ubuntu	20	0	3127984	2.757g	11384	R	99.3	0.1	540:08.20	Ray
25171	ubuntu	20	0	3127996	2.742g	11380	R	99.3	0.1	540:06.75	Ray
25172	ubuntu	20	0	3124056	2.742g	11372	R	99.3	0.1	540:01.96	Ray
25173	ubuntu	20	0	3127692	2.748g	11376	R	99.3	0.1	540:02.28	Ray
25175	ubuntu	20	0	3127996	2.740g	11388	R	99.3	0.1	540:11.21	Ray
25178	ubuntu	20	0	3127676	2.741g	11372	R	99.3	0.1	539:55.19	Ray

This beginning to matter in a lot of new places
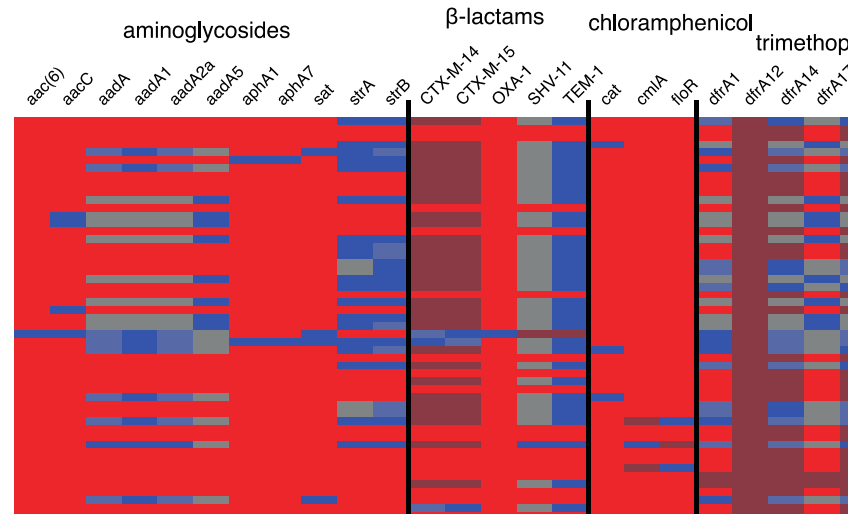

Target Identification
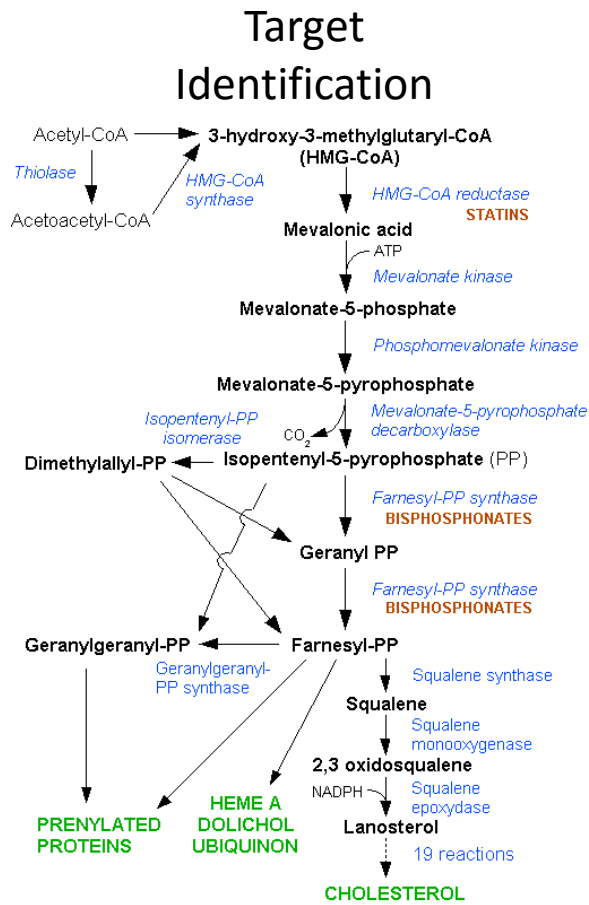
Drug Development
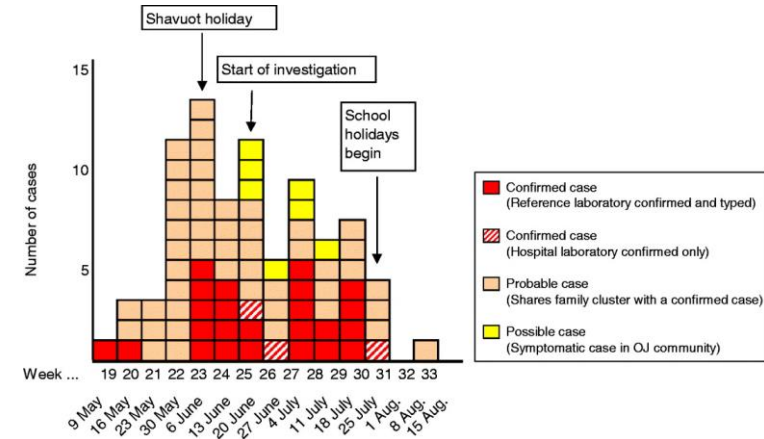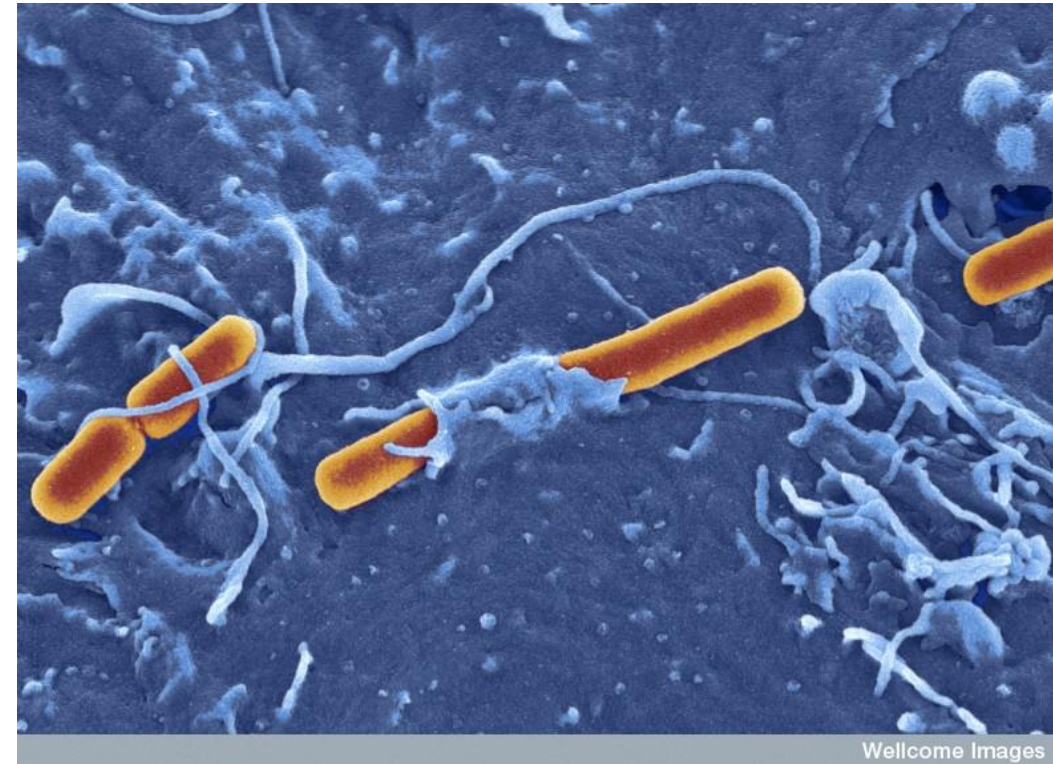
Public Health

Diagnostics

Treatment Selection

Data, Network, Compute

- "Genomics" is more than a single workload

- Most of genomics workloads may not be traditional HPC

- Flexibility is critical
  - Likelihood is that along a single pipeline different elements will have different bottlenecks

- The one missing part in biology is that we must also be **very** user-centric

- In an ideal world, the infrastructure is transparent; it just works, and the users don't even know what it is that they are using

- If HPC is complex, or too hard to use, no matter how perfect, PIs will grow their own

Wellcome Images

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

MRC | Cloud Infrastructure for Microbial Bioinformatics

These Challenges are not only found inside the datacentre

Data generation Multiple TB, every few days

Transport of data from lab

Processing, analysis storage and sharing of data

Size of data generated; Storage and transfer limitations

Volume of data generated; network speed

Data complexity; processing power

The implications

Wellcome Images
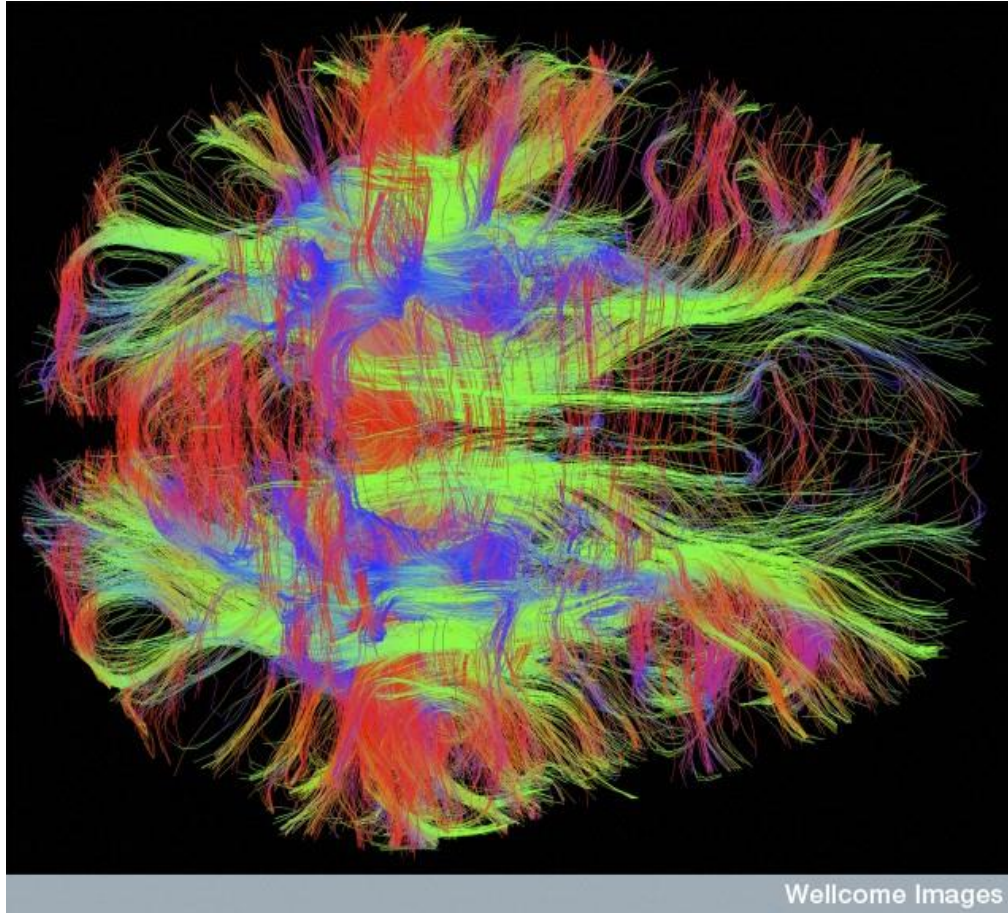
- HPC systems can't be islands
- Need to be part of an integrated infrastructure
- Need to think about ways to more intelligently deliver services
- Move beyond the datacentre to look at solutions that can help with our problems across the data generation chain
- Clearly data and networks are the limiting factor outside the datacentre
- BUT: cleverer compute solutions can alleviate these difficulties

Looking forwards

Use converged architectures to process data close to the sequencing instruments

Move away from dedicated HPC towards a more flexible software defined system

Look for new ways to use new tools as they arise

Cloud Launch

Cloud Launch

Congratulations, your scalable analys

131.251.130.120/galaxy

Galaxy / mGVL 0.10-2   Analyze Data   Workflow   Shared Data   Visualization   Help   User

Tools

search tools

BASIC TOOLS
Get Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Operate on Genomic Intervals
Extract Features

STATISTICS AND VISUALISATION
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data

NGS COMMON TOOLSETS

Welcome to Galaxy on the Cloud
managed by CloudMan

Microbial Genomics Virtual Lab Galaxy

History

search datasets

Unnamed history
0 bytes

This history is empty. You can load your own data or get data from an external source

Using 0 bytes

131.251.130.120

GVL Dashboard   Home   Admin   About

GVL 4.0.0

Welcome to the GVL Dashboard! The GVL Dashboard is a portal through which you can access all services on your GVL instance.

Instance Services

for testcluster

	Service Name	Description	Status	Access Link
	Galaxy	Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.	✓	http://131.251.130.120/galaxy **Username:** manual sign up **Password:** \<custom password\>
	Cloudman	CloudMan is a cloud manager that orchestrates the steps required to provision and manage compute clusters on cloud infrastructure. Use Cloudman to start and manage your Galaxy service and to add additional nodes to your compute cluster.	✓	http://131.251.130.120/cloud **Username:** ubuntu **Password:** \<cluster password\>
	Lubuntu Desktop	Lubuntu is a lightweight desktop environment through which you can run desktop applications on your virtual machine. You can also access the GVL commandline utilities through the desktop.	✓	http://131.251.130.120/vnc **Username:** ubuntu **Password:** \<cluster password\>
	SSH	You can login to your virtual machine remotely through an SSH client.	✓	`ssh ubuntu@131.251.130.120` **Username:** ubuntu **Password:** \<cluster password\>
	JupyterHub	JupyterHub can be used to access your personal IPython Notebook. IPython Notebook is a web-based interactive computational environment where you can combine code execution, text, mathematics, plots and rich media into a single document.	✓	http://131.251.130.120/jupyter **Username:** researcher **Password:** \<cluster password\>
	RStudio	RStudio IDE is a powerful and productive user interface for R.	✓	http://131.251.130.120/rstudio **Username:** researcher **Password:** \<cluster password\>
	Public HTML	This is a shared web-accessible folder. Any files you place in this directory will be publicly accessible.	✓	http://131.251.130.120/public/researcher/ **Username:** **Password:**

131.251.130.120/jupyter/user/ubuntu/notebooks/Untitled.ipynb?kernel_name=python2

jupyter   Untitled   Last Checkpoint: a few seconds ago (unsaved changes)

File   Edit   View   Insert   Cell   Kernel   Help

Code   CellToolbar

In [ ]:

131.251.130.120/rstudio/

R   File   Edit   Code   View   Plots   Session   Build   Debug   Tools   Help

Console ~/

```
R version 3.2.2 (2015-08-14) -- "Fire Safety"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

WARNING: Your CRAN mirror is set to "http://cran.r-project.org" which has an insecure (non-HTTPS)
URL. The repository was likely specified in .Rprofile or Rprofile.site so if you wish to change it
you may need to edit one of those files. You should either switch to a repository that supports HT
TPS or change your RStudio options to not require HTTPS downloads.
```

Environment   History

Import Dataset

Global Environment

Files   Plots   Packages   Help

New Folder   Upload   Delete

Home

Name
galaxy_genomes
public_html

Biological reality

- We don't need an LHC to generate TB's of data a week
- We do this routinely in a (my) lab
- Poses serious (local) computational challenges
- Forces us to consider better how we design our (global) infrastructure
- Users need different things (both in terms of UI, and compute/storage/etc)
- Users also don't want to have to care about what the hardware is underneath
- Typically the user issues are likely to be the main limitation
- The technical stuff is easy; the human interaction stuff is hard

Wellcome Images

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

MRC | Cloud Infrastructure for Microbial Bioinformatics

Conclusions

- Data, compute and storage are all critical
- Need to think beyond the datacentre
- Software is key
  - Need to consider the users in everything ; different users need different things
  - Software makes commodity hardware flexible
  - Flexibility is easier with software

Bottom line: We need to consider how HPC is presented to biologists and what we mean by HPC because if we get it right, we save lives

Wellcome Images
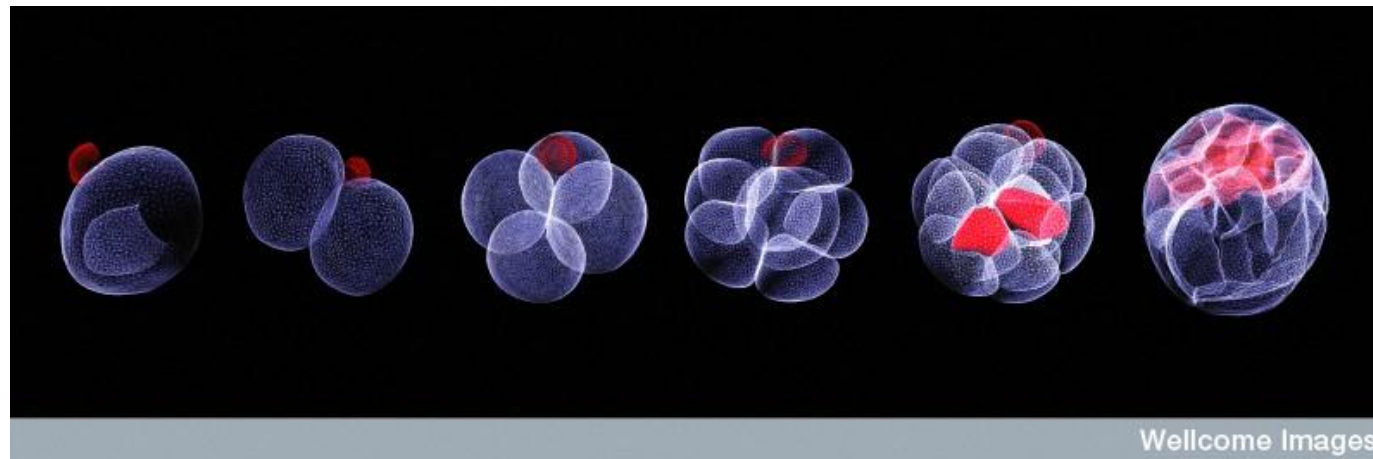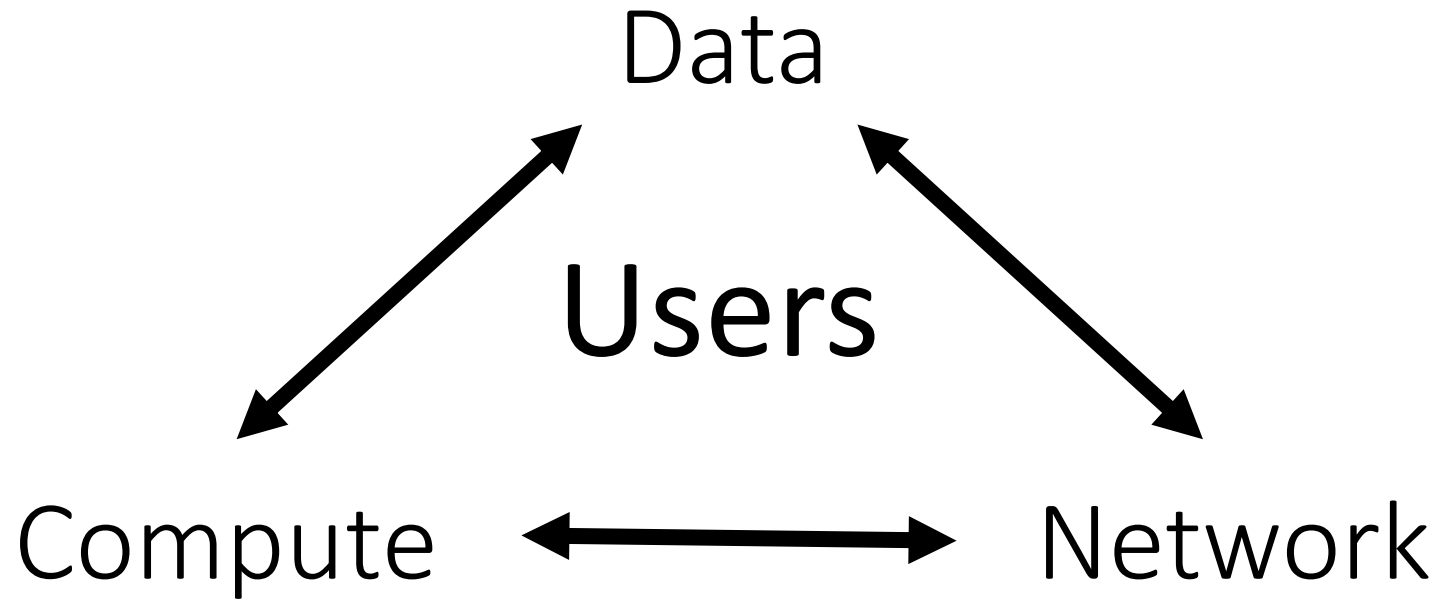
A customers view
Dr Thomas R Connor
School of Biosciences, Cardiff University
connortr@cardiff.ac.uk ; @tomrconnor

Transport

- For many universities and hospitals, the local network will be slow (100Mbit)
- To transfer one run from a high end sequencing instrument could take 4-5 days over this network
- A sequencing run takes ~3 days
- Even 1Gbit would take 10+ hours at line speed
- Replacing this infrastructure is difficult
- Also problematic as issues can occur during transit

Processing, storage and sharing

- Workloads are highly varied

- HTC and HPC

- Range of memory needs

- Produce lots of intermediate (and often small files)

- Outside of a focused institute is a nightmare to design a single system to support this

- Outside of a focused institute with a dedicated informatics team, is a nightmare for sysadmins to support

```
top - 06:55:06 up 1 day, 20:57,  1 user,  load average: 80.85, 81.31, 81.46
Tasks: 916 total,  82 running, 834 sleeping,  0 stopped,  0 zombie
%Cpu(s): 42.3 us, 42.0 sy,  0.0 ni, 15.6 id,  0.0 wa,  0.1 hi,  0.0 si,  0.0 st
KiB Mem:  29230940+total, 18189089+used, 11041852+free,    102388 buffers
KiB Swap:        0 total,        0 used,        0 free. 92621299+cached Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
13447 ubuntu    20   0  0.637t 0.597t   1020 R  99.6 21.9  10282:37 velvetg
25158 ubuntu    20   0 3127832 2.748g  11372 R  99.6  0.1 540:02.18 Ray
25142 ubuntu    20   0 3127740 2.748g  11356 R  99.3  0.1 539:55.40 Ray
25144 ubuntu    20   0 3127776 2.743g  11372 R  99.3  0.1 540:03.22 Ray
25145 ubuntu    20   0 3127684 2.741g  11364 R  99.3  0.1 540:15.43 Ray
25146 ubuntu    20   0 3127844 2.745g  11376 R  99.3  0.1 540:01.13 Ray
25149 ubuntu    20   0 3127676 2.751g  11384 R  99.3  0.1 539:59.63 Ray
25150 ubuntu    20   0 3128148 2.753g  11360 R  99.3  0.1 539:59.58 Ray
25152 ubuntu    20   0 3124052 2.740g  11364 R  99.3  0.1 539:49.83 Ray
25154 ubuntu    20   0 3127800 2.745g  11360 R  99.3  0.1 540:06.91 Ray
25155 ubuntu    20   0 3127976 2.742g  11380 R  99.3  0.1 540:08.55 Ray
25156 ubuntu    20   0 3127740 2.741g  11372 R  99.3  0.1 540:01.02 Ray
25159 ubuntu    20   0 3124120 2.740g  11380 R  99.3  0.1 540:02.65 Ray
25162 ubuntu    20   0 3127824 2.751g  11380 R  99.3  0.1 539:54.52 Ray
25163 ubuntu    20   0 3127672 2.754g  11364 R  99.3  0.1 539:58.78 Ray
25164 ubuntu    20   0 3127684 2.751g  11392 R  99.3  0.1 540:19.34 Ray
25166 ubuntu    20   0 3127708 2.744g  11388 R  99.3  0.1 539:57.87 Ray
25167 ubuntu    20   0 3127996 2.747g  11372 R  99.3  0.1 540:05.34 Ray
25168 ubuntu    20   0 3127704 2.746g  11396 R  99.3  0.1 540:07.64 Ray
25169 ubuntu    20   0 3127824 2.750g  11364 R  99.3  0.1 540:01.92 Ray
25170 ubuntu    20   0 3127984 2.757g  11384 R  99.3  0.1 540:08.20 Ray
25171 ubuntu    20   0 3127996 2.742g  11380 R  99.3  0.1 540:06.75 Ray
25172 ubuntu    20   0 3124056 2.742g  11372 R  99.3  0.1 540:01.96 Ray
25173 ubuntu    20   0 3127692 2.748g  11376 R  99.3  0.1 540:02.28 Ray
25175 ubuntu    20   0 3127996 2.740g  11388 R  99.3  0.1 540:11.21 Ray
25178 ubuntu    20   0 3127676 2.741g  11372 R  99.3  0.1 539:55.19 Ray
```