



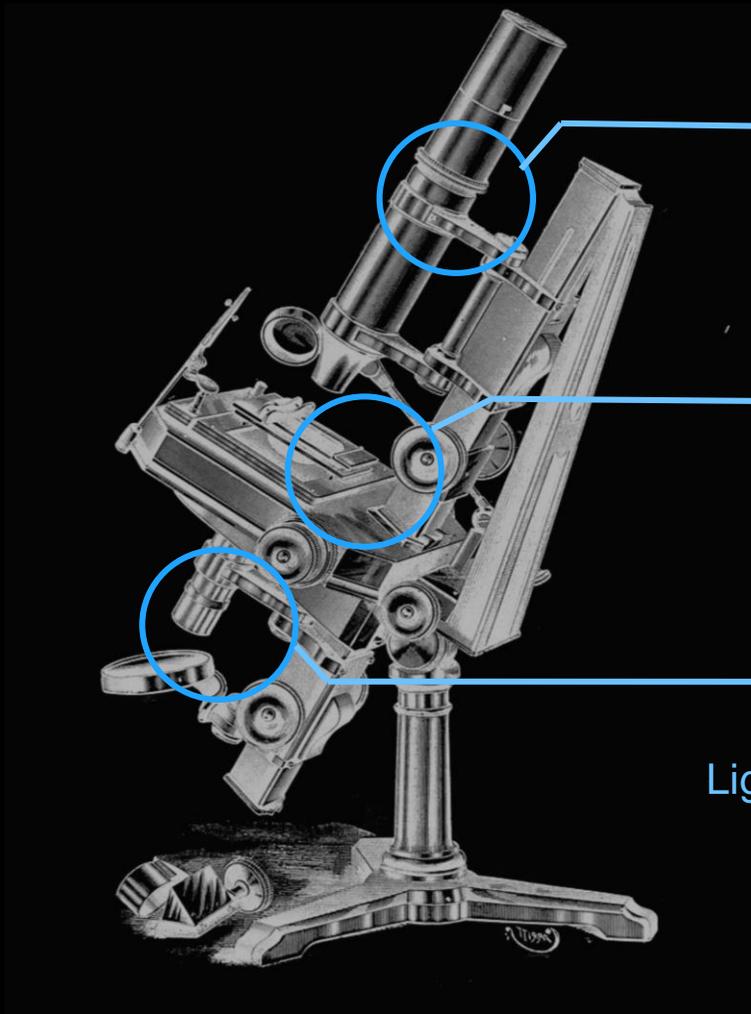
Untapped Data

Initiatives to build ML communities around
big data generators

Dr Jason Rigby

24th Oct 2017

NVIDIA AI Conference, Singapore



INSIGHT
Lens

Visualisation techniques and infrastructure

ANALYSIS
Filters

HPC and cloud Workbenches

CAPTURE
Light Source Samples

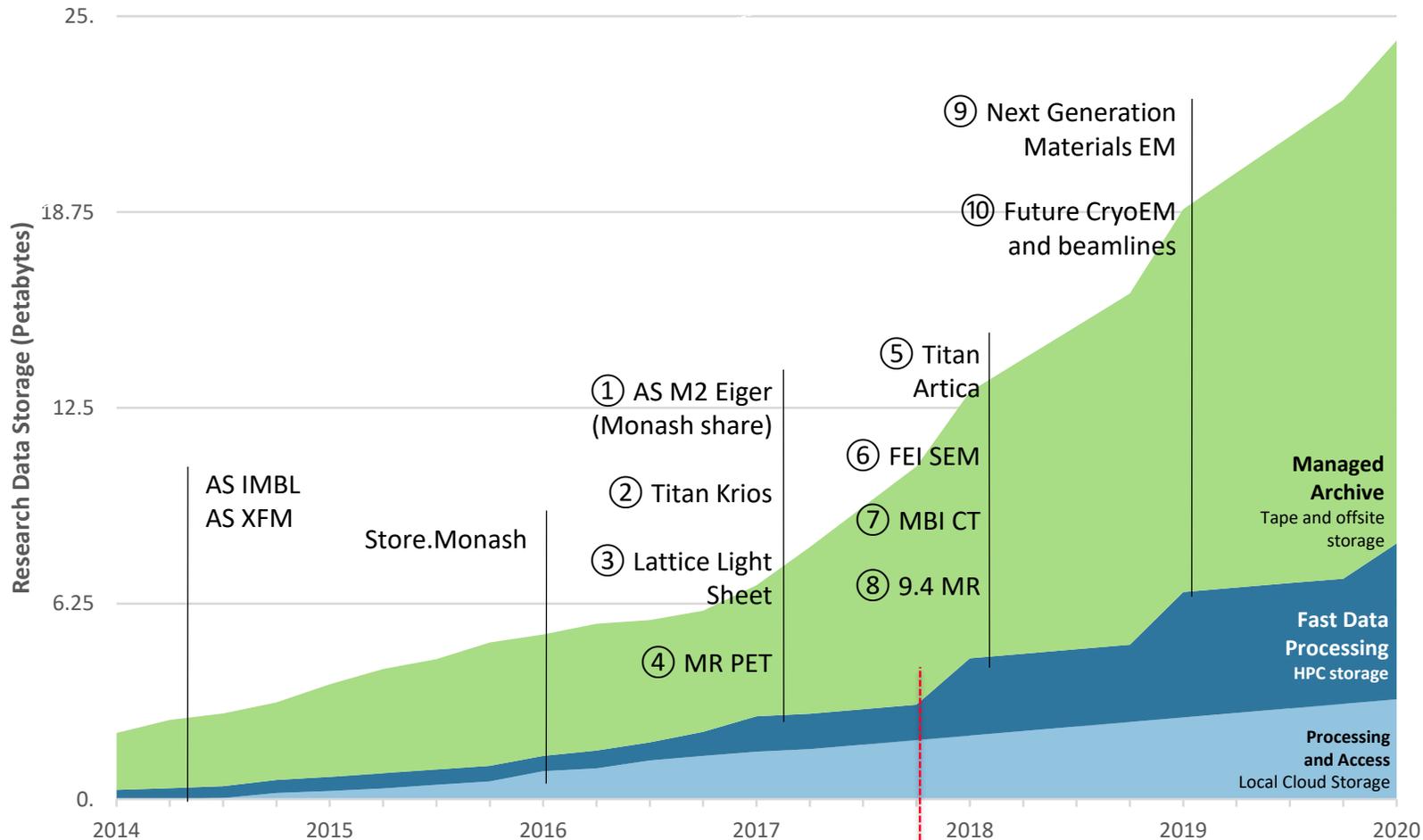
The instruments

The background features a large, dark blue triangle pointing downwards, outlined in white. The interior of the triangle is filled with a dense, intricate pattern of thin, white, overlapping lines that create a sense of depth and complexity. The overall aesthetic is modern and abstract.

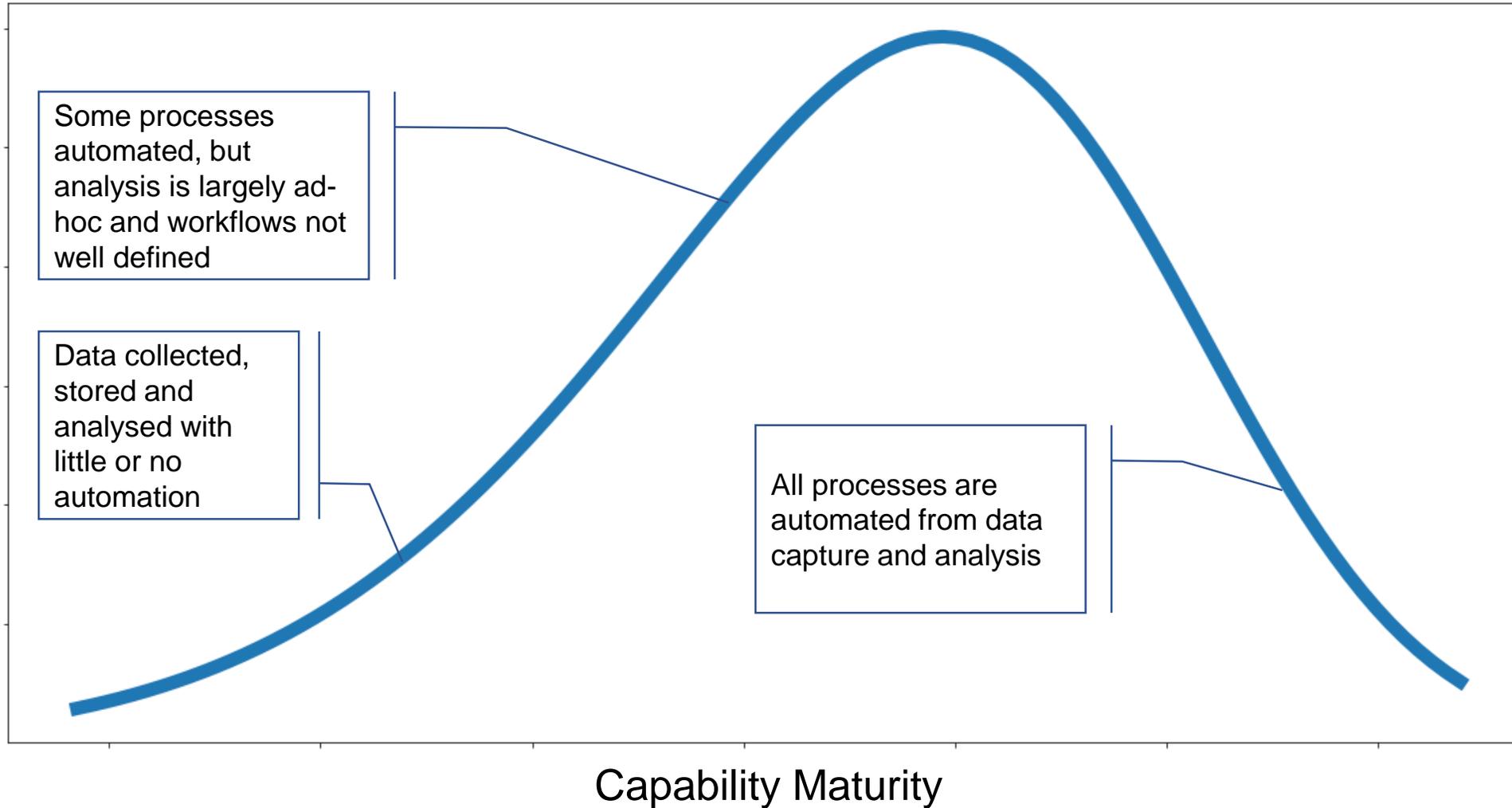
Awareness
**Understanding the
potential**

The data deluge

Data ingestion and storage at Monash University

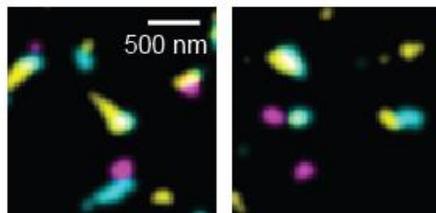
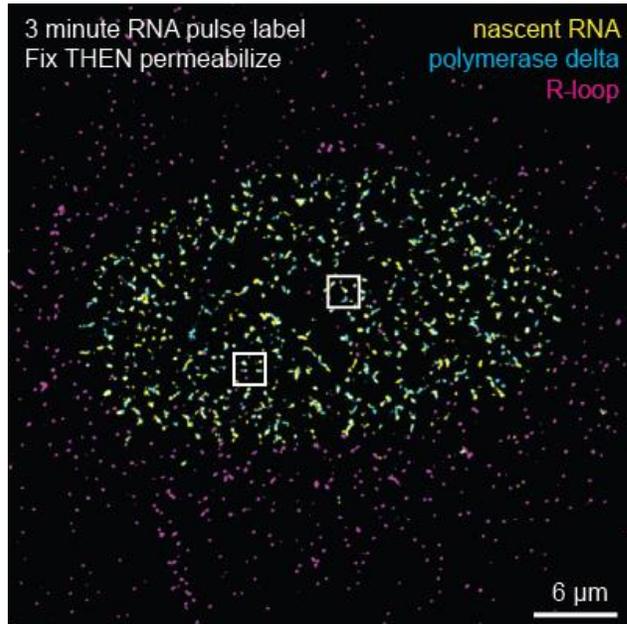


Targeting the long tail

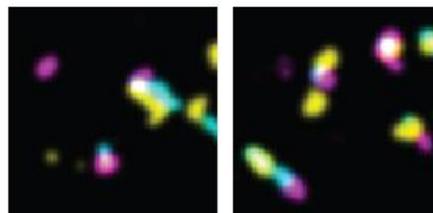
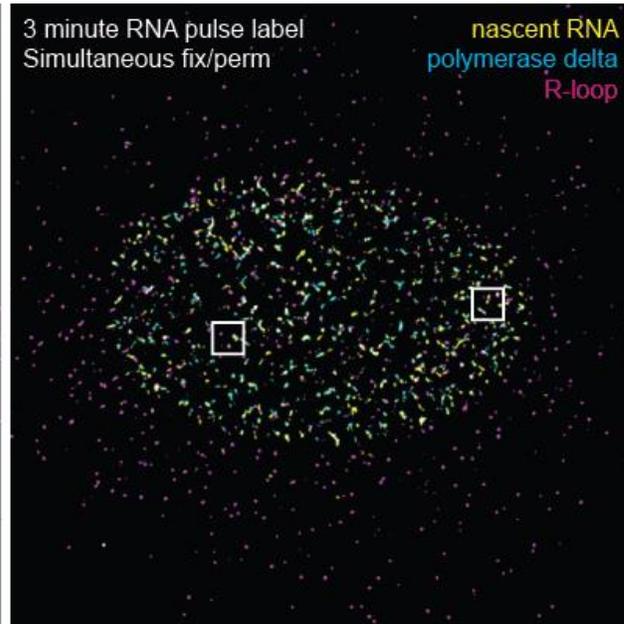


Researchers know their data

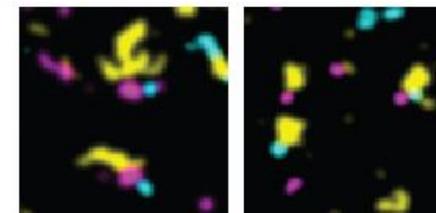
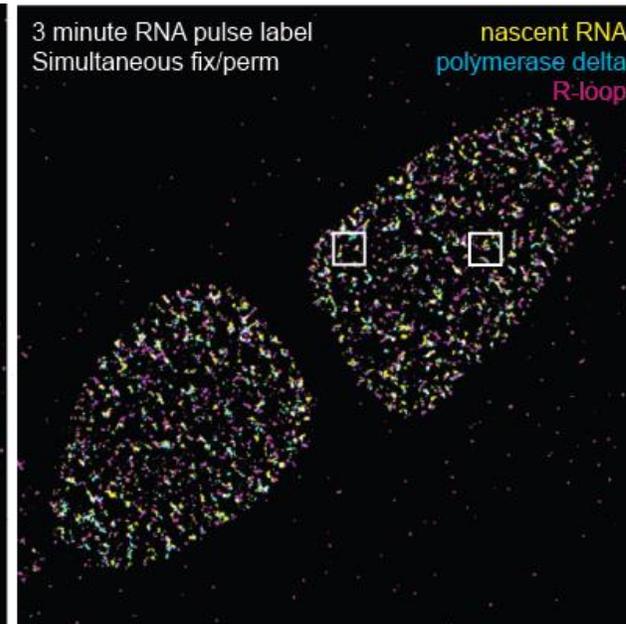
and are best placed to analyse it



4% PFA 37°C 10 mins
1% Triton 37°C 5 mins



4% PFA and 0.5% Triton
37°C 5 mins



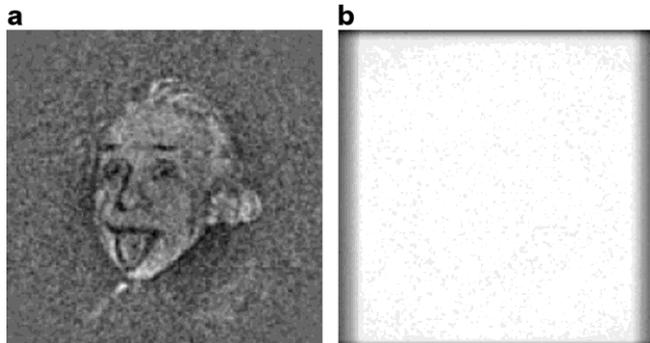
1% Triton 37°C 5 mins
4% PFA 37°C 10 mins

Courtesy of Dr Donna Whelan,
Monash University

Which image is correct?

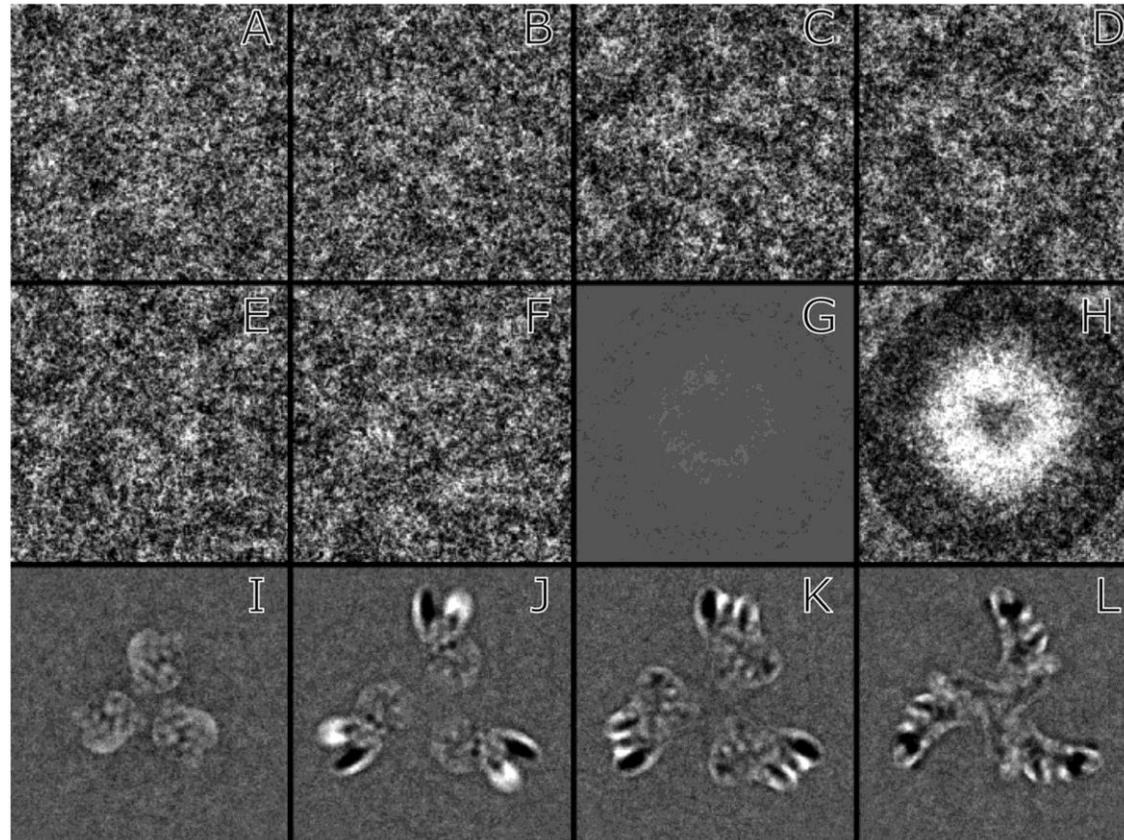
Researchers know their data

and are best placed to analyse it



Shatsky, Maxim, et al. *Journal of structural biology* 166.1 (2009): 67-78.

Is your data noise or the real deal?



Richard Henderson PNAS 2013;110:18037-18041

Researchers know their data

and are best placed to analyse it

Exclusion criteria:

- A history of a diagnosed CVD event defined as myocardial infarction (MI), heart failure, angina pectoris, stroke, transient ischemic attack, >50% carotid stenosis or previous carotid endarterectomy or stenting, coronary artery angioplasty or stenting, coronary artery bypass grafting, or abdominal aortic aneurysm;
 - A clinical diagnosis of atrial fibrillation;
 - Serious illness likely to cause death within the next 5 years;
 - A current or recurrent condition with a high risk of major bleeding;
 - Anemia (hemoglobin <12 g/dl males <11 g/dl females);
 - An absolute contraindication or allergy to aspirin;
 - Current participation in an ongoing clinical trial;
 - Current use of aspirin for secondary prevention;
 - Current continuous use of other antiplatelet drug or anticoagulant;
 - A systolic blood pressure ≥ 180 mm Hg and/or a diastolic blood pressure ≥ 105 mm Hg;
- ASPREE Investigator Group. *Contemporary clinical trials* 36.2 (2013): 555-564.

Population	Non -Anaemia*
Children 6 - 59 months of age	110 or higher
Children 5 - 11 years of age	115 or higher
Children 12 - 14 years of age	120 or higher
Non-pregnant women (15 years of age and above)	120 or higher
Pregnant women	110 or higher
Men (15 years of age and above)	130 or higher

± Adapted from references 5 and 6

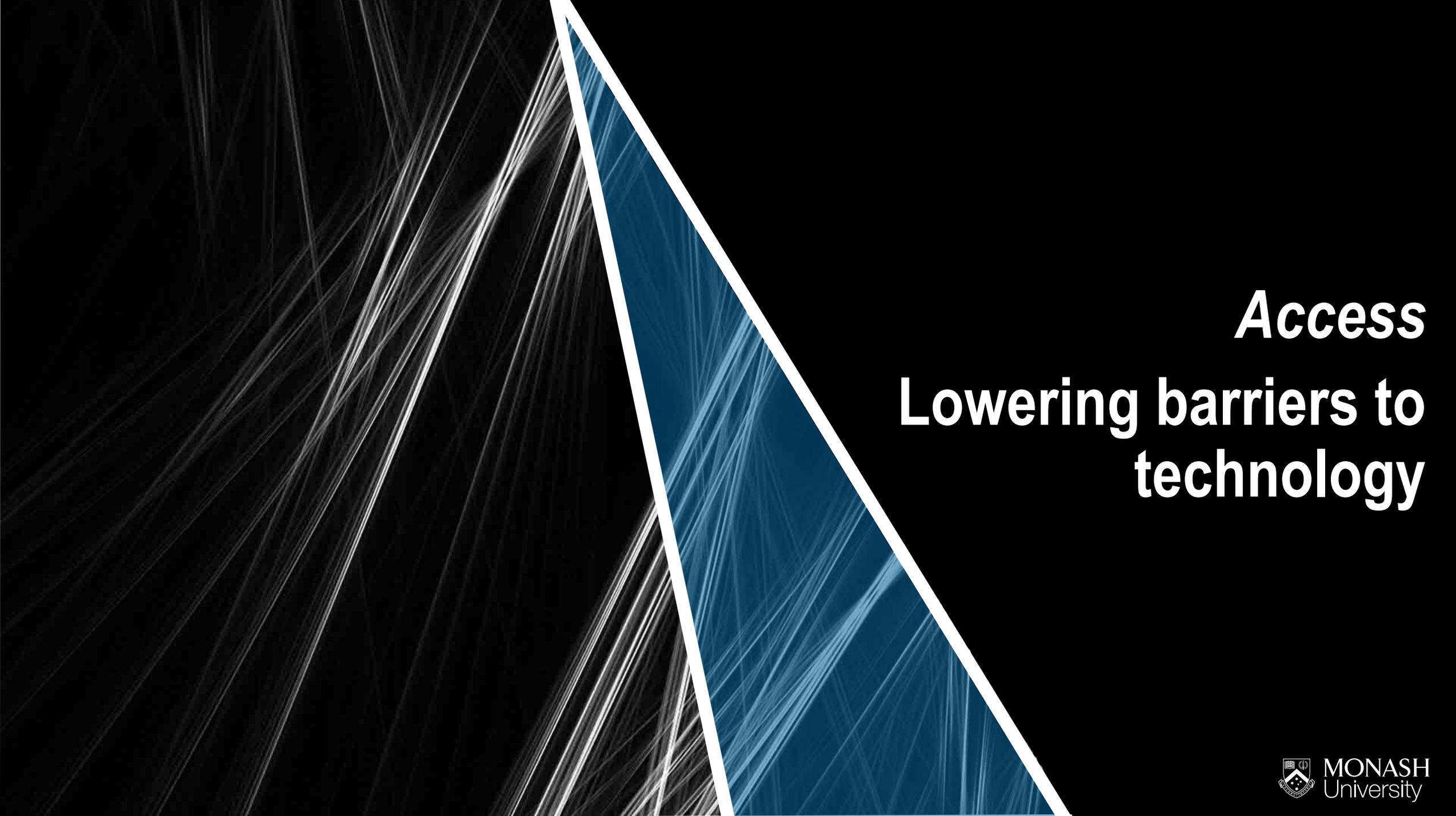
* Haemoglobin in grams per litre

WHO criteria for anemia

Can you use the data for your intended purpose?

With great data comes great responsibility.

Researchers must be included in ML advancements to drive innovation in their discipline.

The background features a large, dark blue triangle pointing downwards, outlined in white. The interior of the triangle is filled with a dense, intricate pattern of thin, white, overlapping lines that create a sense of depth and complexity. The overall aesthetic is modern and technological.

Access
**Lowering barriers to
technology**

MASSIVE 3

Stage 2 upgrade

1,600 Intel Haswell CPU-cores

2,448 Intel Skylake CPU-cores

NVIDIA GPU coprocessors for data processing and visualisation:

48 NVIDIA Tesla K80

40 NVIDIA Pascal P100

60 NVIDIA V100

2 NVIDIA DGX1-V

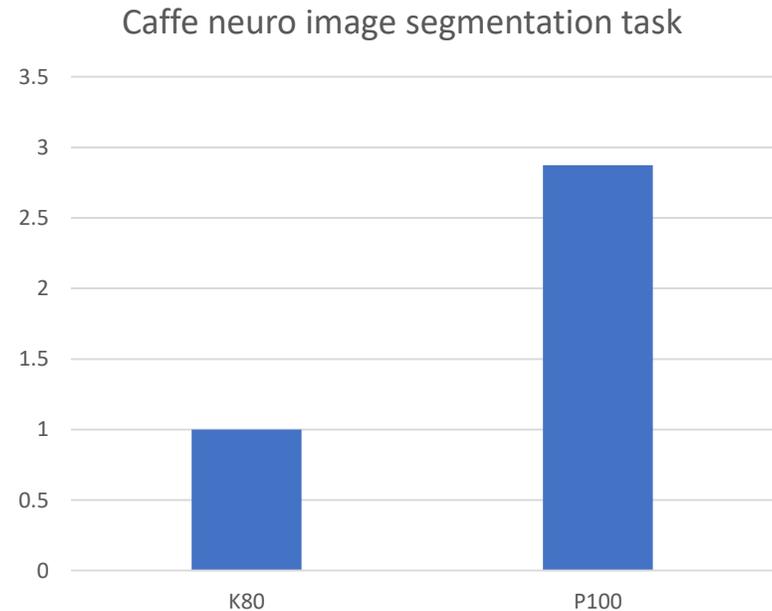
8 NVIDIA Grid K1 GPUs for medium and low end visualisation

A 1.15 petabyte Lustre parallel file system

A 2 petabyte Lustre parallel file system upgrade

100 Gb/s Ethernet Mellanox Spectrum

Supplied by Dell, Mellanox and NVIDIA





 @vintage_computer_museum



NCI supercomputer

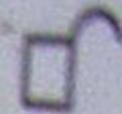
```
[jrigby@m3-login2 ~]$
```

POWER

TURBO

RESET

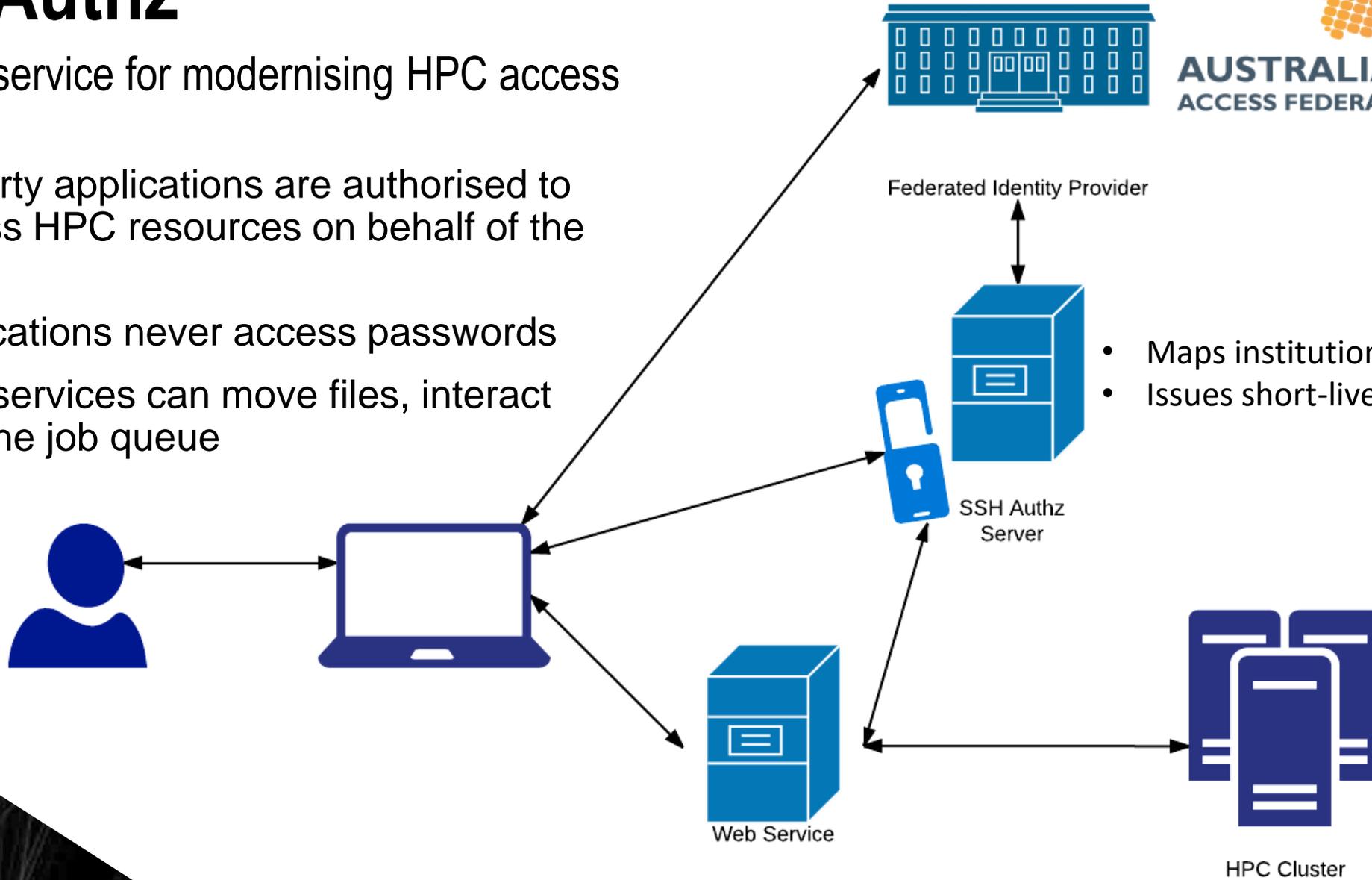
H-DISK



SSH Authz

Our web service for modernising HPC access

- 3rd party applications are authorised to access HPC resources on behalf of the user
- Applications never access passwords
- Web services can move files, interact with the job queue

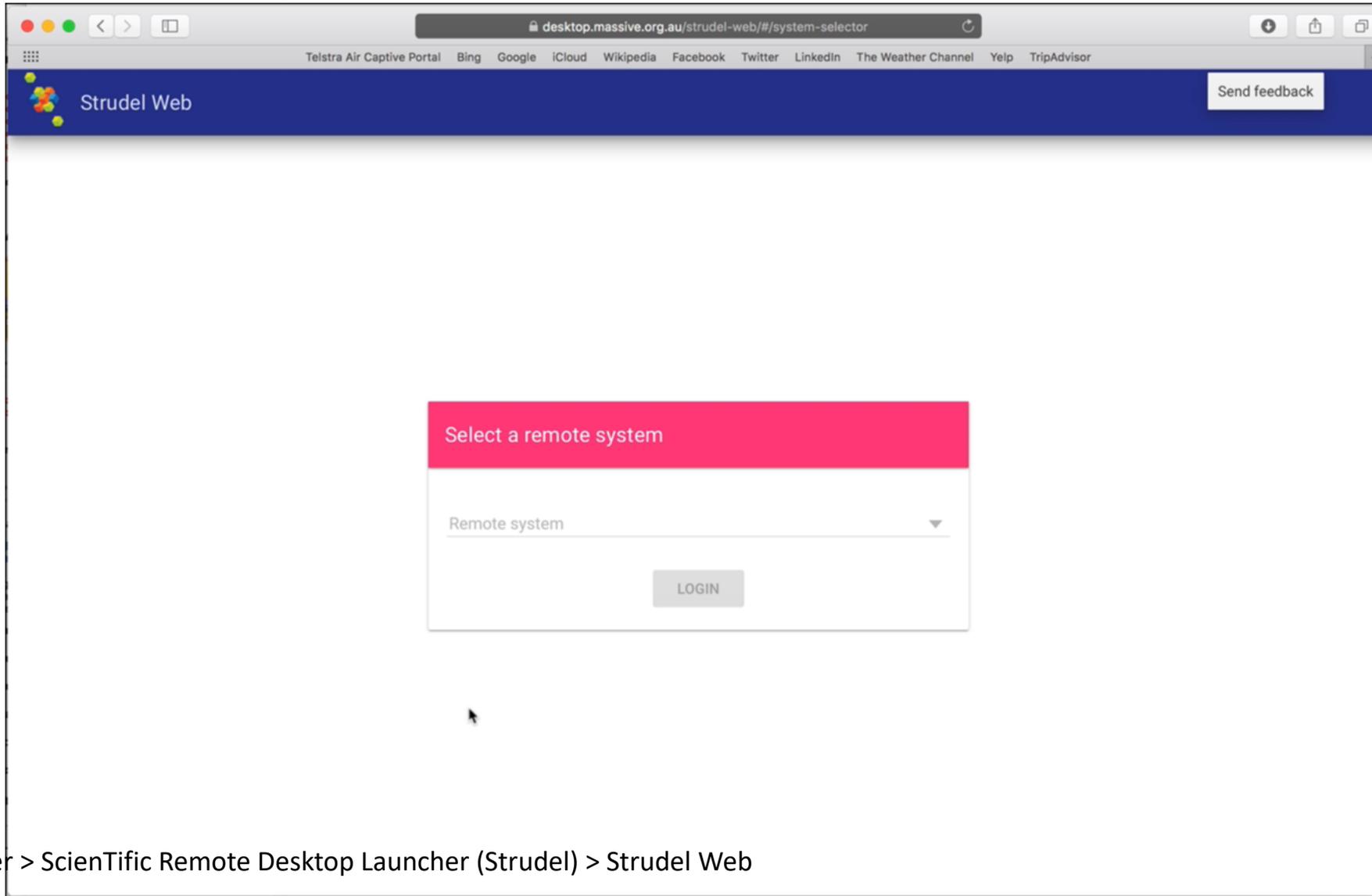


- Maps institutional account to UNIX
- Issues short-lived certificates

<https://github.com/monash-merc/ssh-authz>

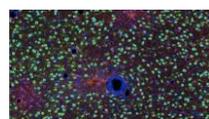
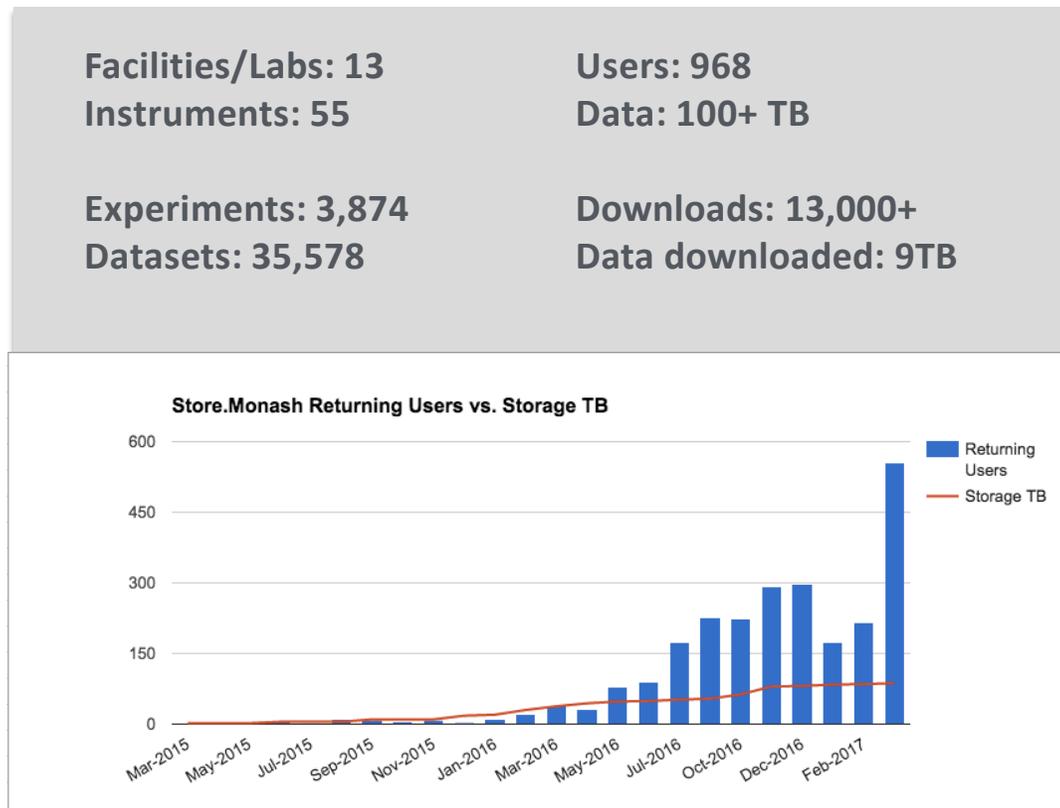
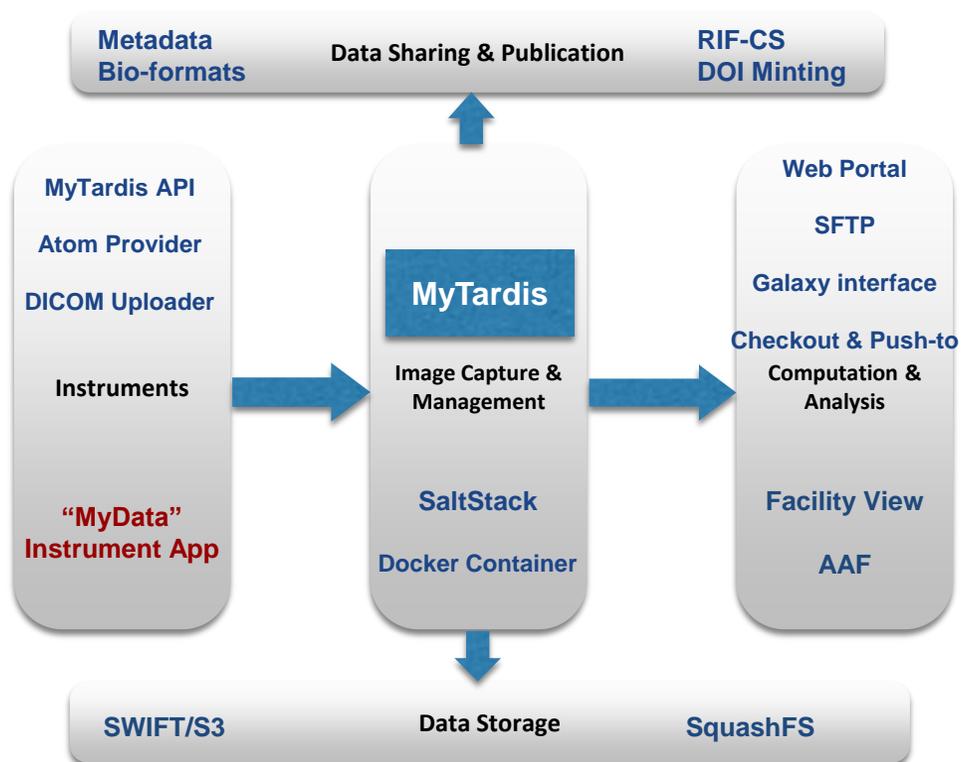
Strudel Web Demo

Running a remote desktop in the browser



Store.Monash

Our institutional instrument data repository



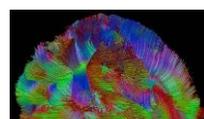
Monash Micro Imaging



CryoEM



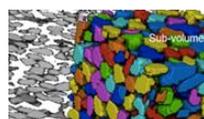
FlowCore



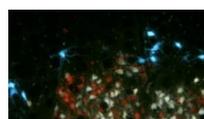
MBI



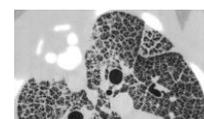
MBPF



XMFIG



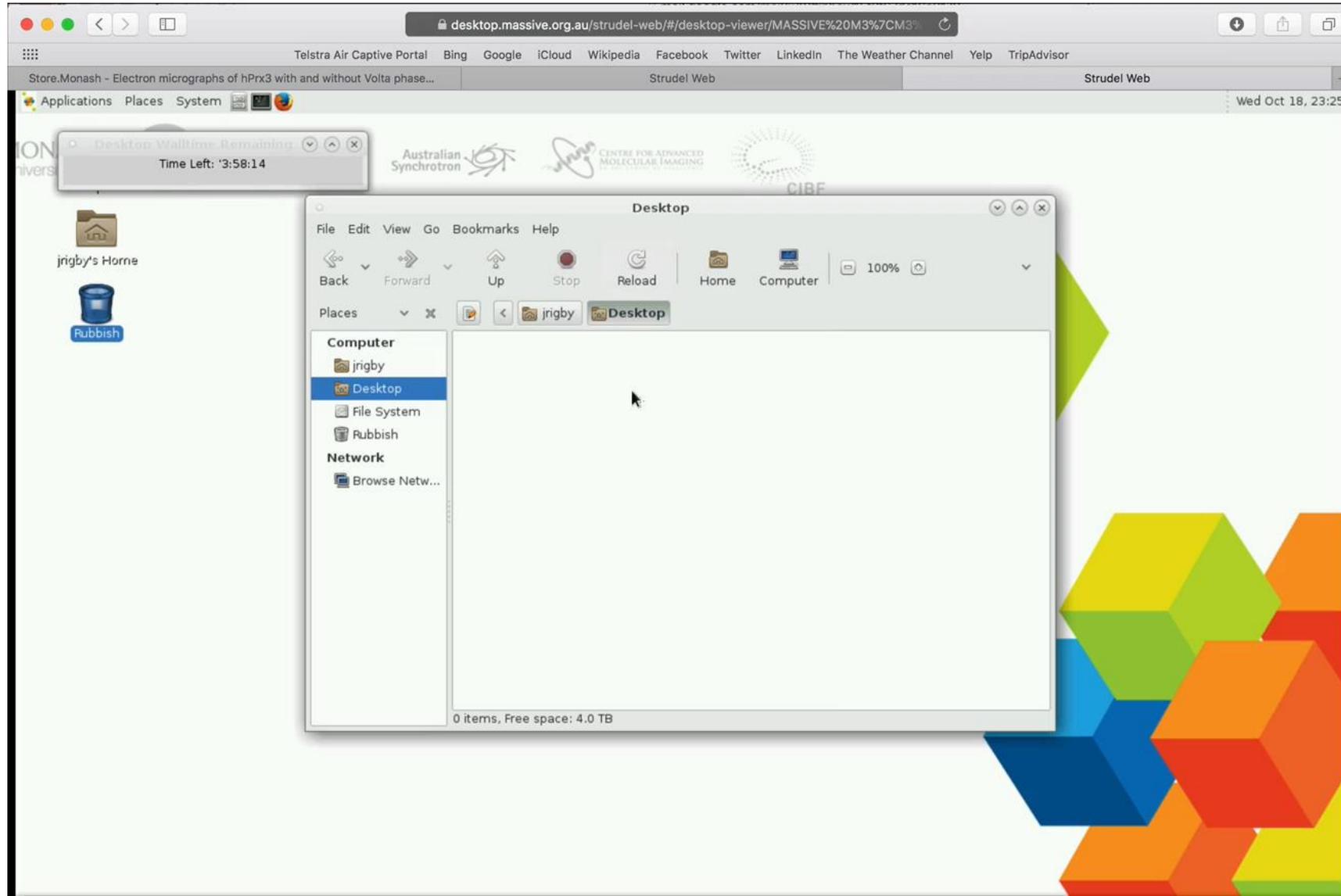
Rosa Lab



Kitchen Lab

Store.Monash Push-To Demo

Moving data from storage to compute



Supporting HPC-based web apps

DIGITS and Jupyter on-demand: Work in progress

The screenshot shows the Strudel Web interface. At the top left, there is a logo and the text "Strudel Web". To the right of this is a "Send feedback" button. Below the header, the user's email "Jason.Rigby@monash.edu" is displayed. The main content area is titled "M3 Standard Desktop". On the left side, there is a sidebar with the heading "MASSIVE M3" and two options: "M3 STANDARD DESKTOP" (selected) and "M3 LARGE DESKTOP". The main content area has a section titled "Launch a desktop" with a dropdown arrow. Below this is a configuration table with columns for "Nodes", "Processors per node", "Memory (gb)", and "Time". The values are 1, 3, 13, and 4 respectively. A "LAUNCH" button is centered below the table. Below the configuration table is a section titled "Running desktops" with a refresh icon. The text below this section reads "You currently have no running desktops." At the bottom of the main content area is a section titled "Server messages" with the text "Any messages from the server will be displayed here".

Strudel Web

Send feedback

Jason.Rigby@monash.edu

M3 Standard Desktop

MASSIVE M3

M3 STANDARD DESKTOP

M3 LARGE DESKTOP

Launch a desktop

Nodes	Processors per node	Memory (gb)	Time
1	3	13	4

LAUNCH

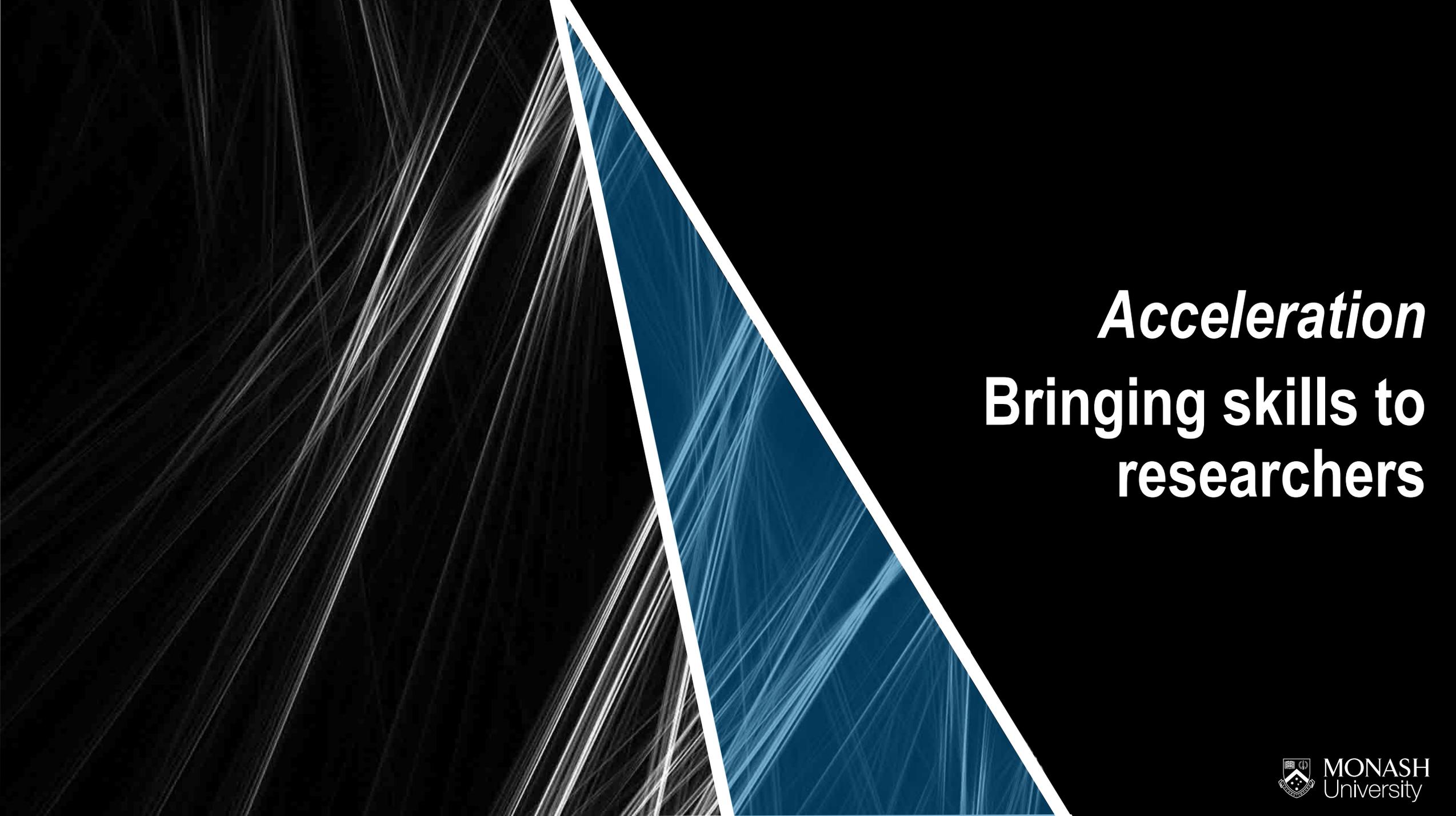
Running desktops

You currently have no running desktops.

Server messages

Any messages from the server will be displayed here





Acceleration
**Bringing skills to
researchers**

**Enabling ML capabilities is a
partnership between technology
providers and researchers**

ASpirin in Reducing Events in the Elderly

ASPREE

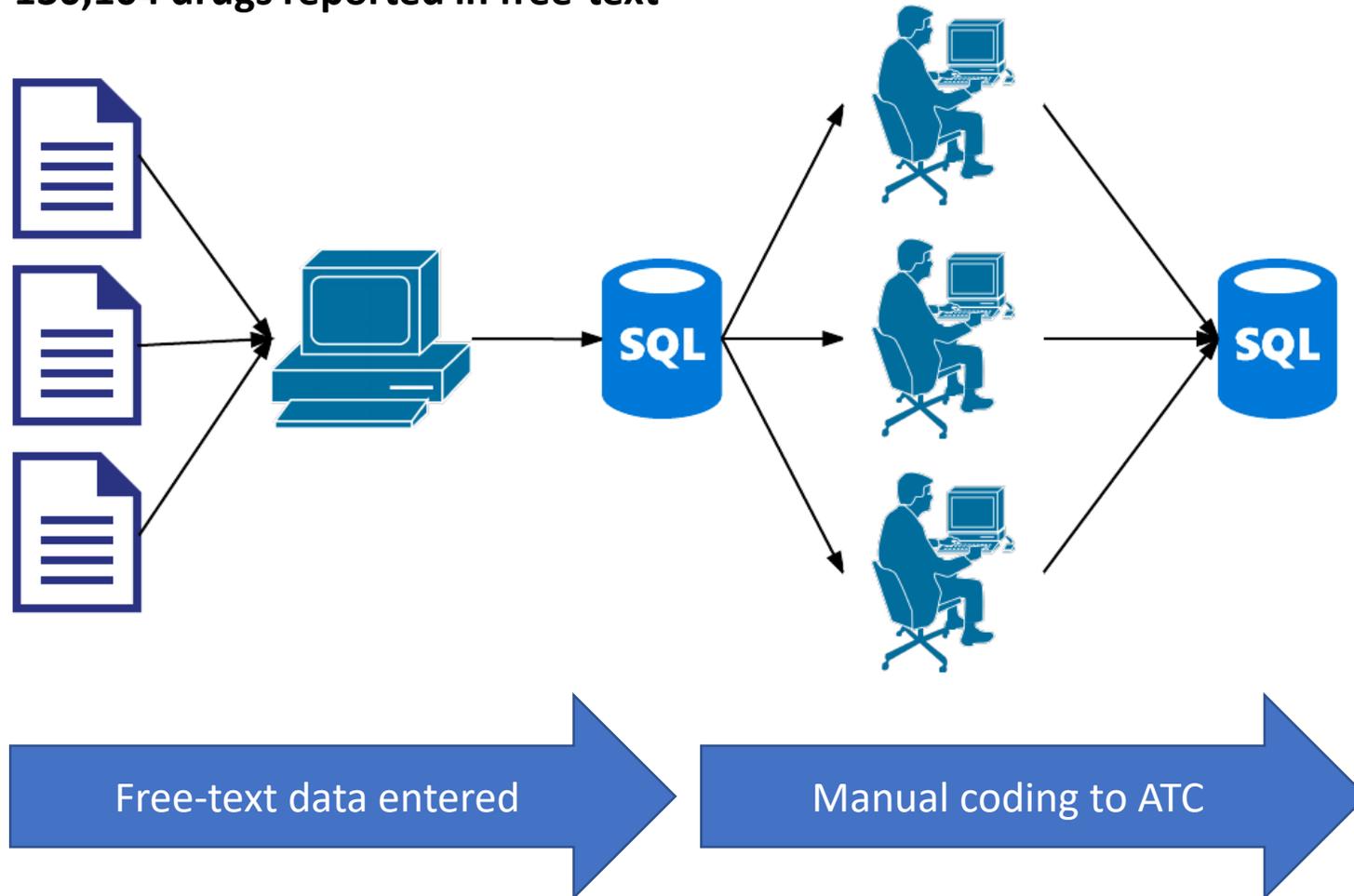


= 19,114

ASPIrin in Redicing Events in the Elderly

Autocoding concomitant medications

130,104 drugs reported in free-text



ASpirin in Redicing Events in the Elderly

Autocoding concomitant medications

A many:one mapping problem – drug names can be:

- Generic
- Trade names
- Misspelt generics
- Misspelt trade names
- Free-text with additional information (e.g. “... taken 3x daily”)
- Differ between countries

Many can be autocoded with
an RNN text classifier!



**ATC Code:
N02BE01**

ASPIrin in Redicing Events in the Elderly

Autocoding concomitant medications

ASPREE ConMed Search

Enter a medication name for a list of possible ConMed database matches:

acetaminophen

acetomenefen

ID	Drug Name	Pr
649	PARACETAMOL	0.3005886972
1131	EPLERENONE	0.1904084235
1900	PREDNISOLONE (Oral)	0.0852287039
1901	PREDNISOLONE (Enema)	0.0594153032
2085	TRIAMCINOLONE - TOPICAL	0.0590389036
1962	CAFFEINE, ERGOTAMINE	0.0502391607
554	MOMETASONE	0.0358436182
111	BETAMETHASONE	0.0286783669
394	SOMATROPIN	0.0275046453
2015	HYDROCODONE, PARACETAMOL	0.0256109536

ASPREE ConMed Search

Enter a medication name for a list of possible ConMed database matches:

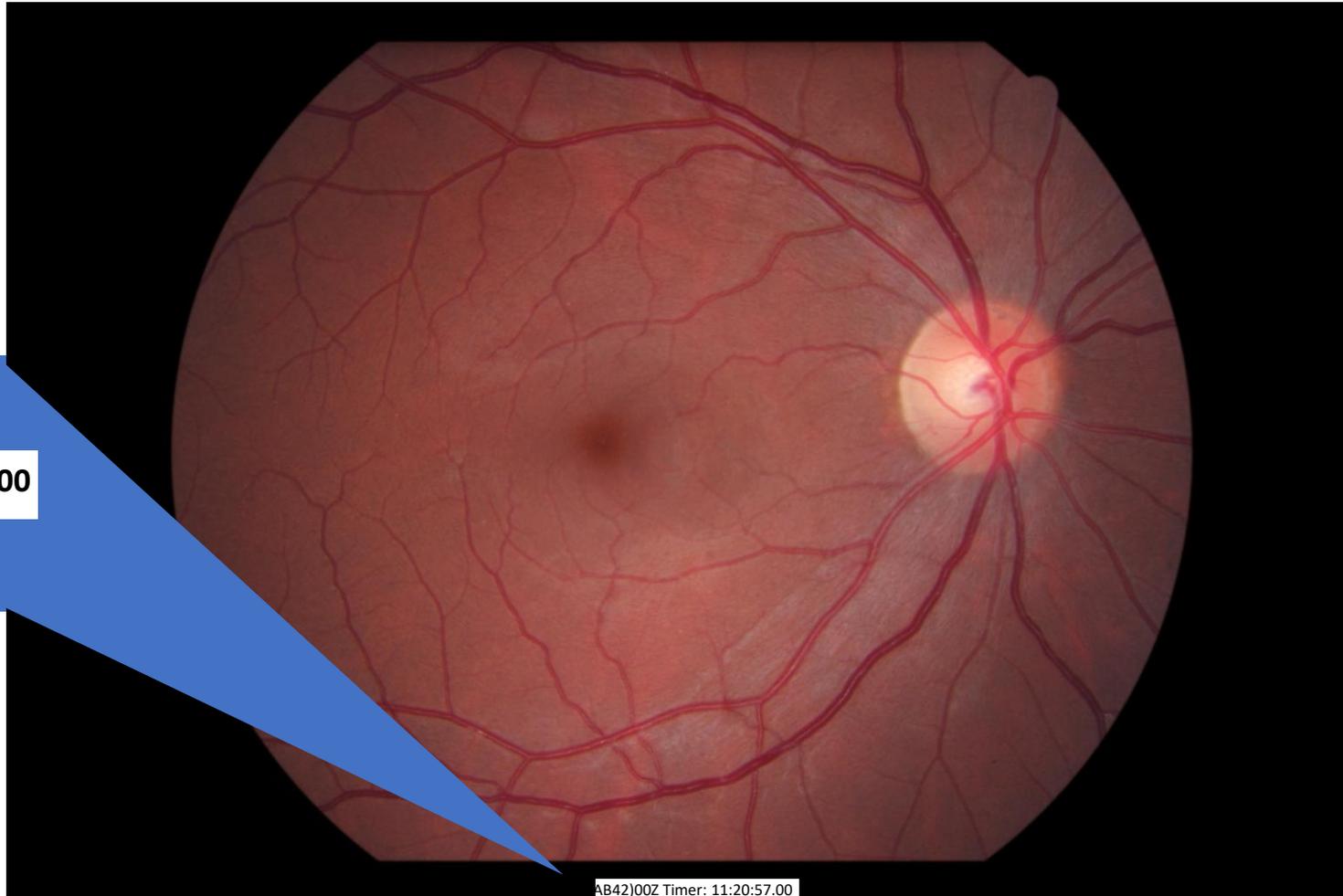
atorvastatin

torvostatin

ID	Drug Name	Pr
86	ATORVASTATIN	0.9807870388
2002	EZETIMIBE - ATORVASTATIN	0.0057576662
745	ROSUVASTATIN	0.0042524473
1935	AMLODIPINE- ATORVASTATIN	0.0039550546
1263	LOVASTATIN	0.0022850621
2125	SODIUM CROMOGLYCATE	0.0015681938
1633	LORSTAT	0.000544793
1608	EZETIMIBE - SIMVASTATIN	0.0002928106
683	PRAVASTATIN	0.000155759
2117	NYSTATIN (TOPICAL)	0.0000666033

ASpirin in Redicing Events in the Elderly

Extracting image identifiers from retinal scans



99% Accuracy reidentifying
mislabelled images

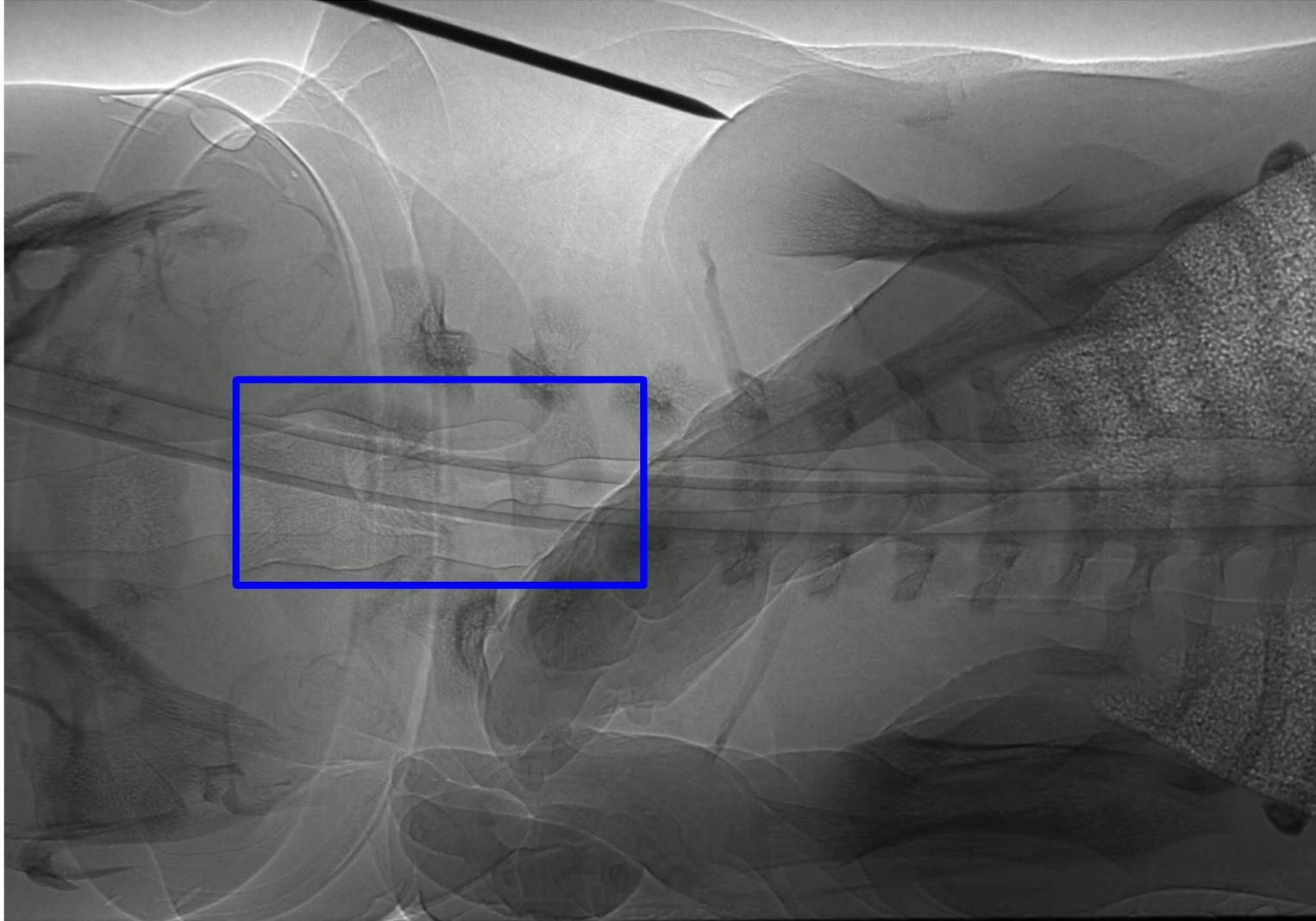
Applied to 7000+ unlabeled
images

AB42)00Z Timer: 11:20:57.00

AB42)00Z Timer: 11:20:57.00

Preterm respiratory development

Identifying the state of the glottis in rabbit kittens



Work in progress:
~70% accuracy in
correctly identifying the
state of the glottis

Courtesy of Dr Marcus
Kitchen & group,
Monash University

Data science toolbox

A containerised stack of curated DL and ML applications

```
302 lines (269 sloc) | 7.99 KB
Raw Blame History
1 Bootstrap: debootstrap
2 MirrorURL: http://us.archive.ubuntu.com/ubuntu/
3 OSVersion: xenial
4 Include: apt
5 %post
6 apt install -y software-properties-common
7 apt-add-repository -y universe
8
9 # Speed up the build
10 export NCPUS=`grep -c ^processor /proc/cpuinfo`
11 export MAKEFLAGS="-j $NCPUS"
12
13 #####
14 # Install CUDA and CUDNN #
15 #####
16 export CUDA_VERSION=8.0.61
17 export CUDNN7_VERSION=7.0.1.13
18 export CUDNN6_VERSION=6.0.21
19 export CUDNN5_VERSION=5.1.10
20 export CUDA_PKG_VERSION="8-0=${CUDA_VERSION}-1"
21 export NVIDIA_GPGKEY_SUM=d1be581509378368edeec8c1eb2958702feedf3bc3d17011adbf24efacce4ab5
22 export NVIDIA_GPGKEY_FPR=ae09fe4bbd223a84b2ccfce3f60f4b3d7fa2af80
23
```



<https://github.com/monash-merc/data-sci-singularity>

Workshops and training

Providing hands-on access to educational materials for upskilling



OCT.
30

Deep Learning for Life Sciences Workshop

by Monash University and the NVIDIA Deep Learning Institute

Free

Workshops and training

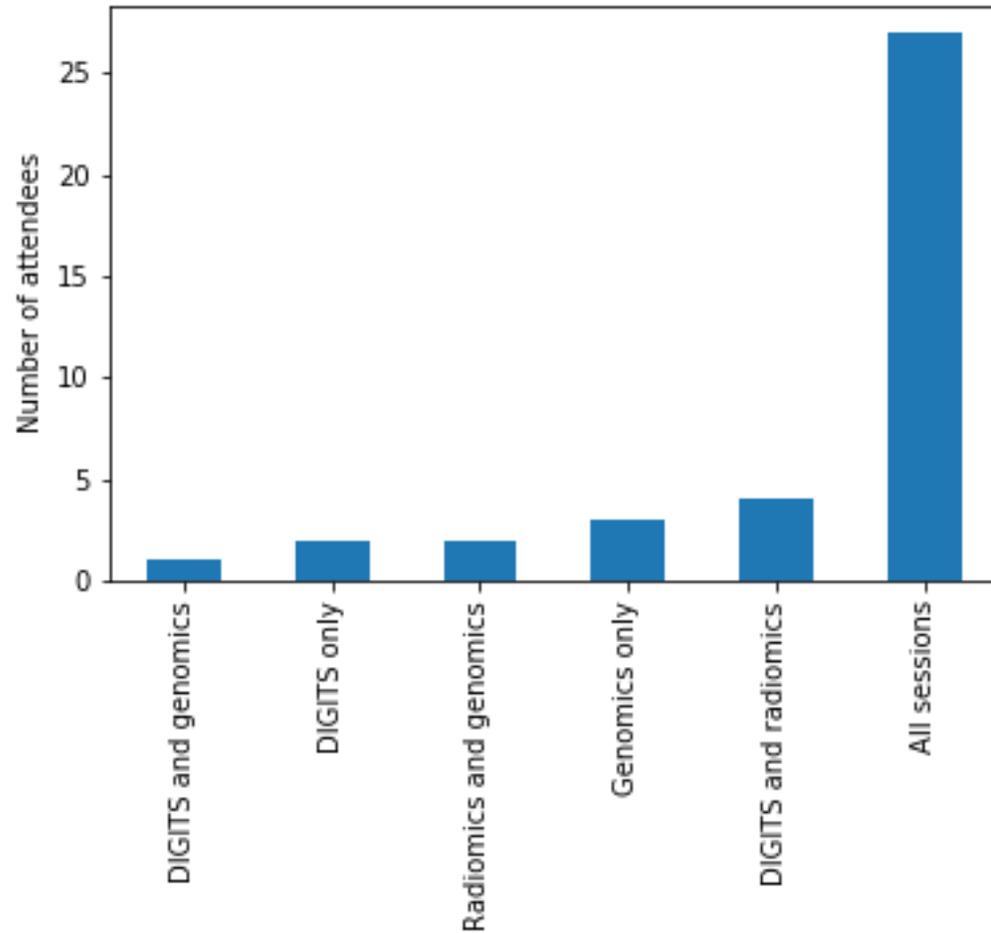
Providing hands-on access to educational materials for upskilling

Sales by Ticket Type

TICKET TYPE	PRICE	SOLD	STATUS	END SALES
Medical Image Segmentation with DIGITS (Includes introduction and CNN discussion)		33/50	On Sale	29/10/17 6:00 pm
Image Classification with TensorFlow		33/50	On Sale	29/10/17 6:00 pm
Deep Learning for Genomics using DragoNN with Keras and Theano		34/50	On Sale	29/10/17 6:00 pm

Workshops and training

Providing hands-on access to educational materials for upskilling



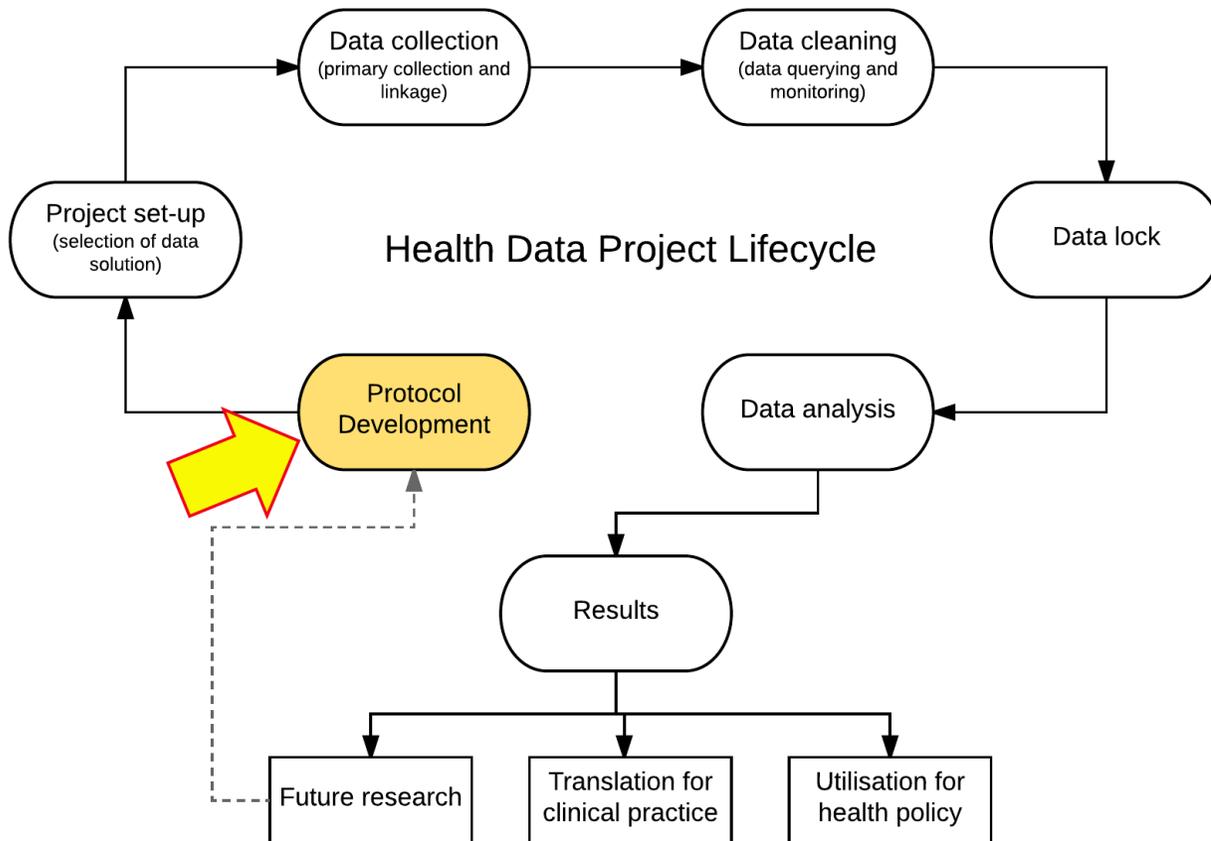
Attendees include:

- Monash staff and students
- Neighboring universities
- Government
- Other non-university research institutions

Workshops and training

Modernising the delivery of data-centric curricula

Master of Public Health: Practical Data Management



jupyter

Control Panel

Logout

Files

Running

Clusters

Select items to perform actions on them.

Upload

New



	Name ↑	Last Modified ↑
<input type="checkbox"/>	data	16 days ago
<input type="checkbox"/>	ClearHead Data Preparation.ipynb	2 months ago
<input type="checkbox"/>	ClearHead Data Quality Reporting-solutions.ipynb	17 days ago
<input type="checkbox"/>	ClearHead Data Quality Reporting.ipynb	17 days ago
<input type="checkbox"/>	Week 12 - Outcome of ClearHead.ipynb	16 days ago
<input type="checkbox"/>	Week 2 - Introduction to Jupyter and R (Part 1).ipynb	3 months ago
<input type="checkbox"/>	Week 3 - Introduction to Jupyter and R (Part 2).ipynb	3 months ago
<input type="checkbox"/>	Week 5 - Merging and Reshaping-Copy1.ipynb	2 months ago
<input type="checkbox"/>	Week 5 - Merging and Reshaping.ipynb	2 months ago
<input type="checkbox"/>	Week 8 - Aggregation.ipynb	a month ago

Take-home messages

- The quantity and variety of data is increasing
- The most qualified people to analyse their data using ML techniques should be the researchers themselves
- Cultivating ML communities is a partnership between technology and service providers and the researchers
- Empowerment through outreach and training is essential to raise awareness and generate ideas
- Ease of access will increase technology uptake amongst the “long tail”