

Automating Data Visualization



December 2018

Nick Holliman @Binocularity

Professor of Visualization, Newcastle University
Fellow of The Alan Turing Institute, London

The Alan Turing Institute

The national institute for data science and
artificial intelligence

Building on a strong scientific legacy



- Alan Turing's pioneering work in theoretical and applied mathematics, engineering and computing are considered to be the key disciplines comprising the field of data science.
- *"I propose to consider the question, "Can machines think?"..."*

In 1950 Turing published his seminal paper, *Computing Machinery and Intelligence*, which is credited with laying the foundations for the development and philosophy of artificial intelligence.

Founding the Institute

“We will found The Alan Turing Institute to ensure Britain leads the way again in the use of big data and algorithm research”

Chancellor of the Exchequer

Budget Speech, March 2014

**The
Alan Turing
Institute**

EPSRC

Engineering and Physical Sciences
Research Council

Network of industry,
charity, government
partners

Network of
university
members

Strategic
government
investment

The goals of the Institute

Innovate and develop world-class research in data science and artificial intelligence

Apply our data science research to real-world problems, supporting the creation of new products, services, and jobs

Train the next generation of data science and artificial intelligence leaders

Advising policy-makers and shaping the public conversation around data

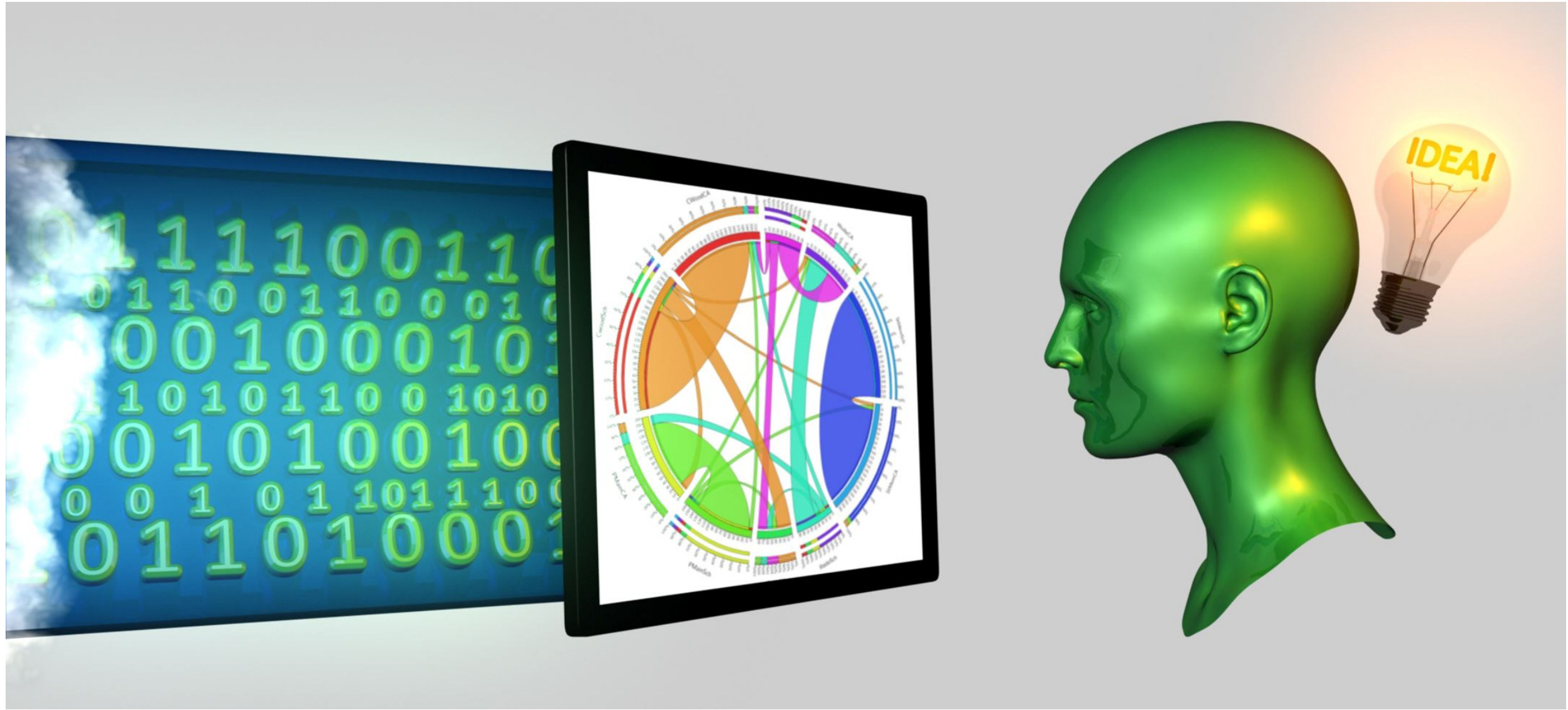
Our university network



The Institute's partners and collaborators



Why do we need Visualization?

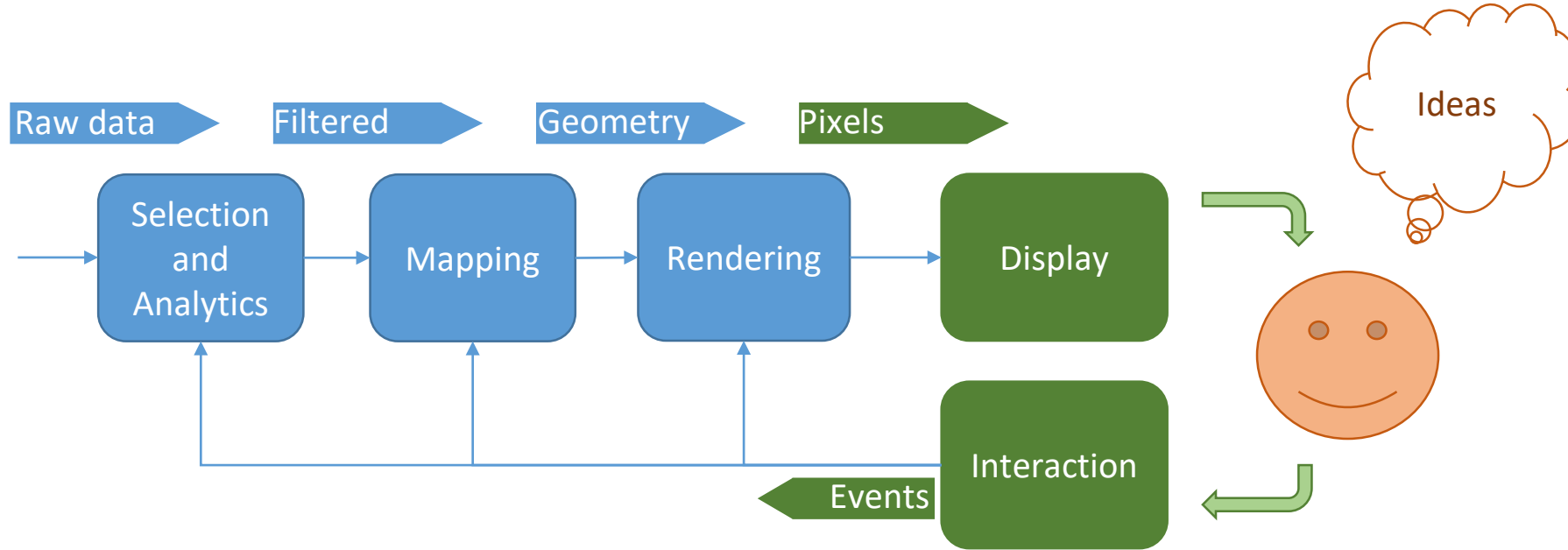


Big-data is getting bigger: humans still need to be able to access, use and compare data.

Scalable visualization using the cloud

“Scalable real-time visualization using the cloud”
Holliman and Watson, IEEE Cloud Computing, Dec 2015

Classical visualization pipeline

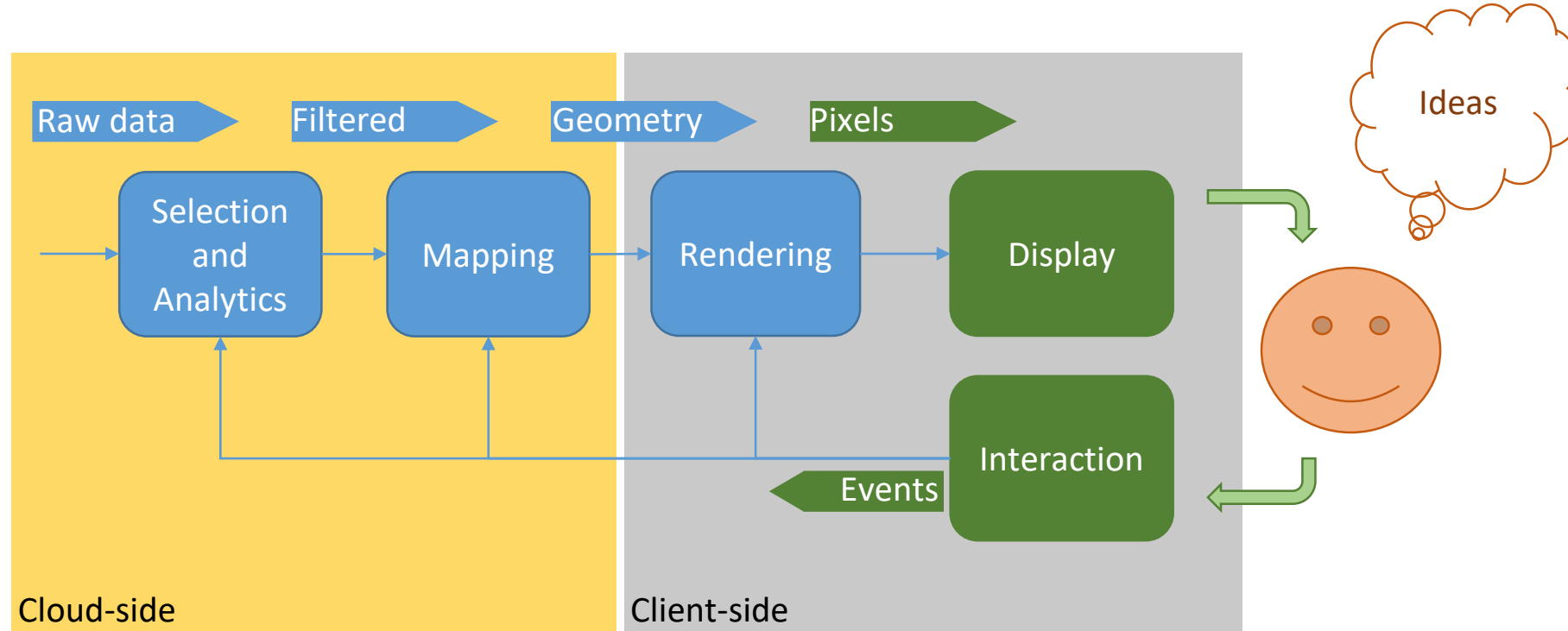


A data flow model of how information is converted via visual computing to images matched to human understanding.

This will often include significant interactive elements.

How can we map this to the cloud for big-data problems?

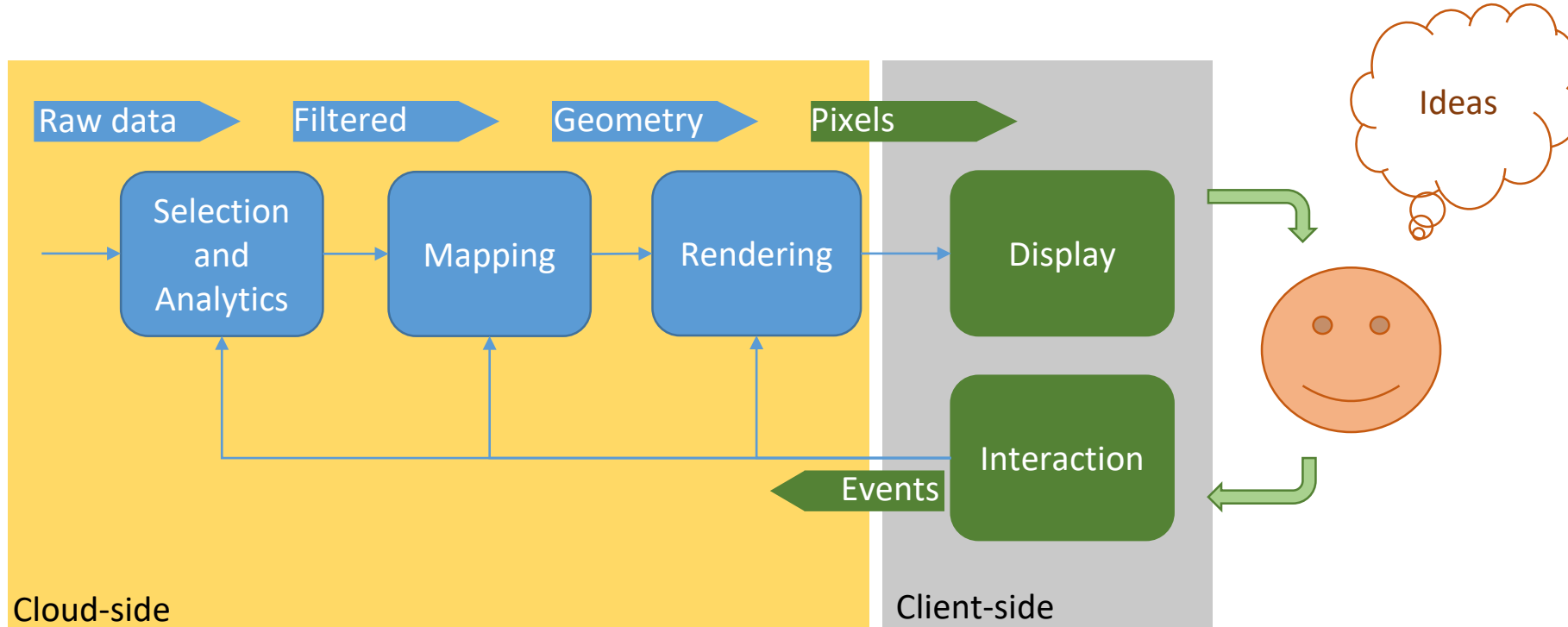
Client-side rendering (send-geometry)



This splits the pipeline between mapping and rendering: model data (eg SVG 2D or UML 3D) that is transmitted from the cloud to the client.

The recent trend has been to load the client (web browser) with complex rendering tasks.

Cloud-side rendering (send-image)



For big data client side rendering is a potential bottleneck.

An alternative is to split the pipeline between rendering and display:
now it is the image (pixels) that are transmitted from the cloud to the client.

Client-side vs Cloud-side rendering

	Client-side send-geometry	Cloud-side send-image
Bandwidth (to client)	Unbounded must handle whole model changes in worst case	Bounded by human spatial/temporal acuity (4K needs ~ 25Mbps)
Bandwidth (to cloud)	All interactive event traffic.	All interactive event traffic.
Latency	Delays need to be masked by caching the model up to cache limit	Required to be low enough to support interaction (50-100mS)
Client complexity	Variable functionality graphics support needed plus caching.	Fixed function image display client, graphics power in the cloud.

#TeraScope: automating data visualization

Establish the feasibility of scalable cloud supercomputing

Mark Turner (Newcastle, Computing)

Stephen Dowsland (Newcastle, Computing)

James Charleton (Northumbria, VNG)

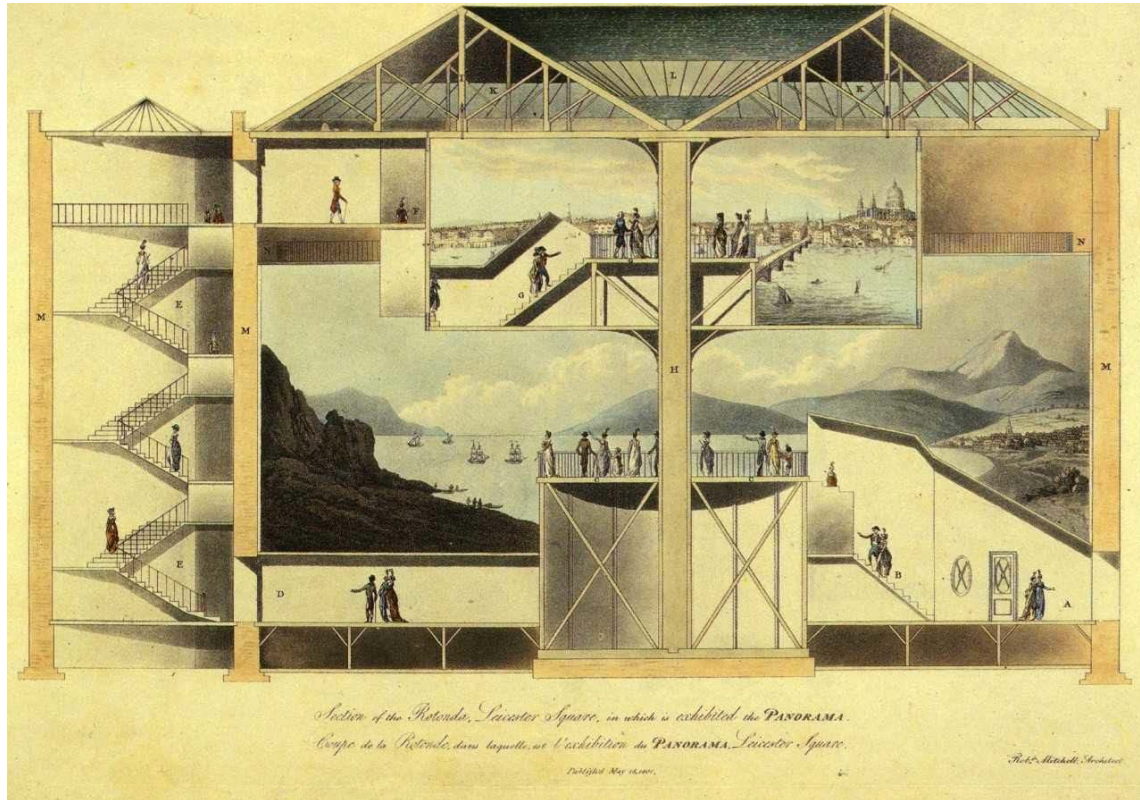
Manu Anthony (Newcastle, Computing)

Phil James (Newcastle, Engineering)

“A scalable platform for visualization using the cloud”
Dowsland, Turner and Holliman, CGAT, April 2017,
Singapore

History of panoramic images for urban data-scapes

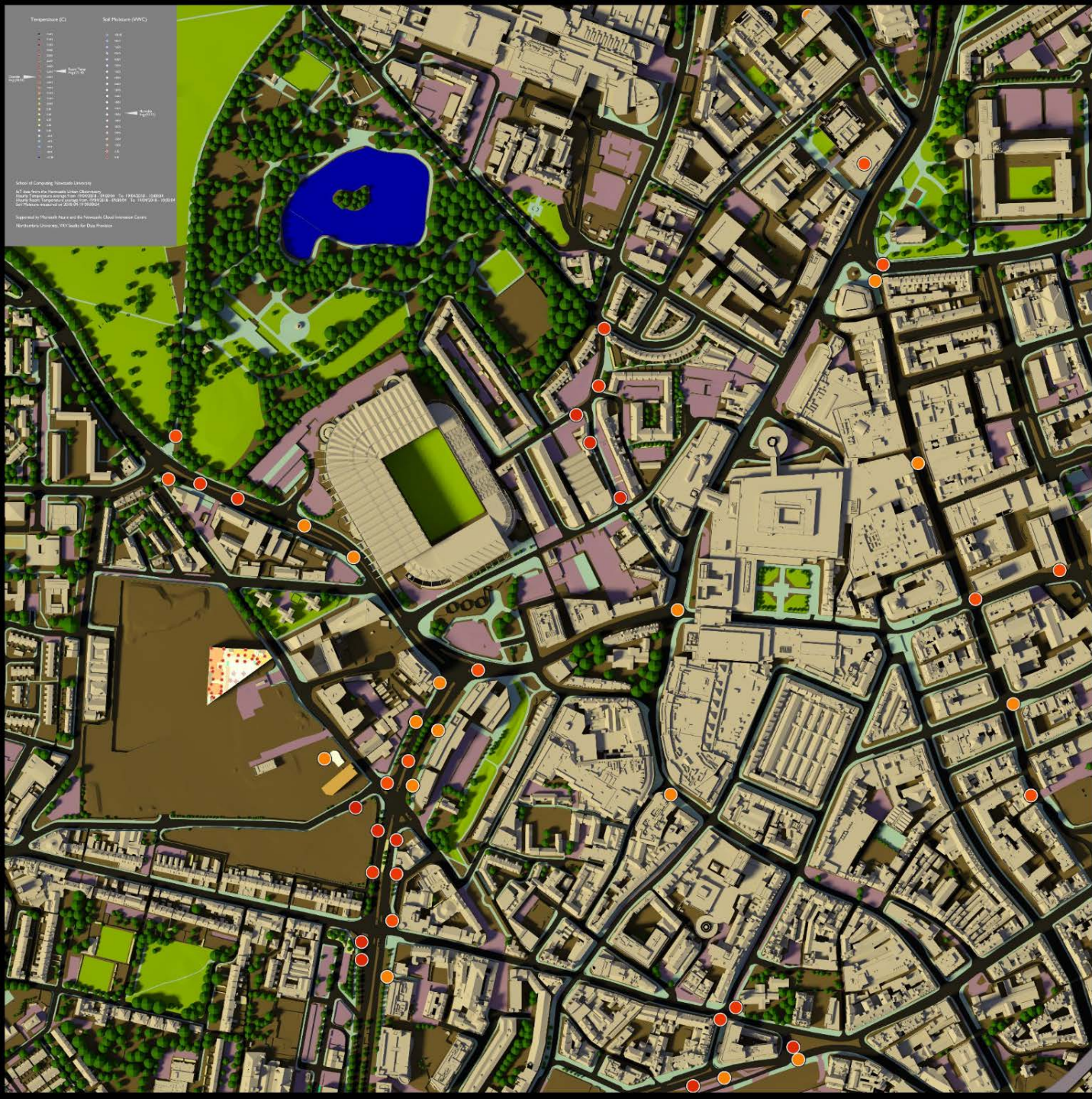
How can we represent urban data so that is widely accessible and contains an ability to be viewed at multiple scales?



Robert Barker, 1787, granted first patent on panoramic imaging, with the aim of immersing visitors in an (urban) scene.



In 1877 Eadweard Muybridge took cityscape panoramas of San Francisco, labelling every building in the key.



#TeraScope

One TeraScope image is:
1,099,511,627,776 pixels

equivalent to:
530,243 full HD TV images

The **krpano** viewer supports
accessible multi-scale viewing
of urban data.

How far can we zoom in?

Whole image:

1.28 km x 1.28 km

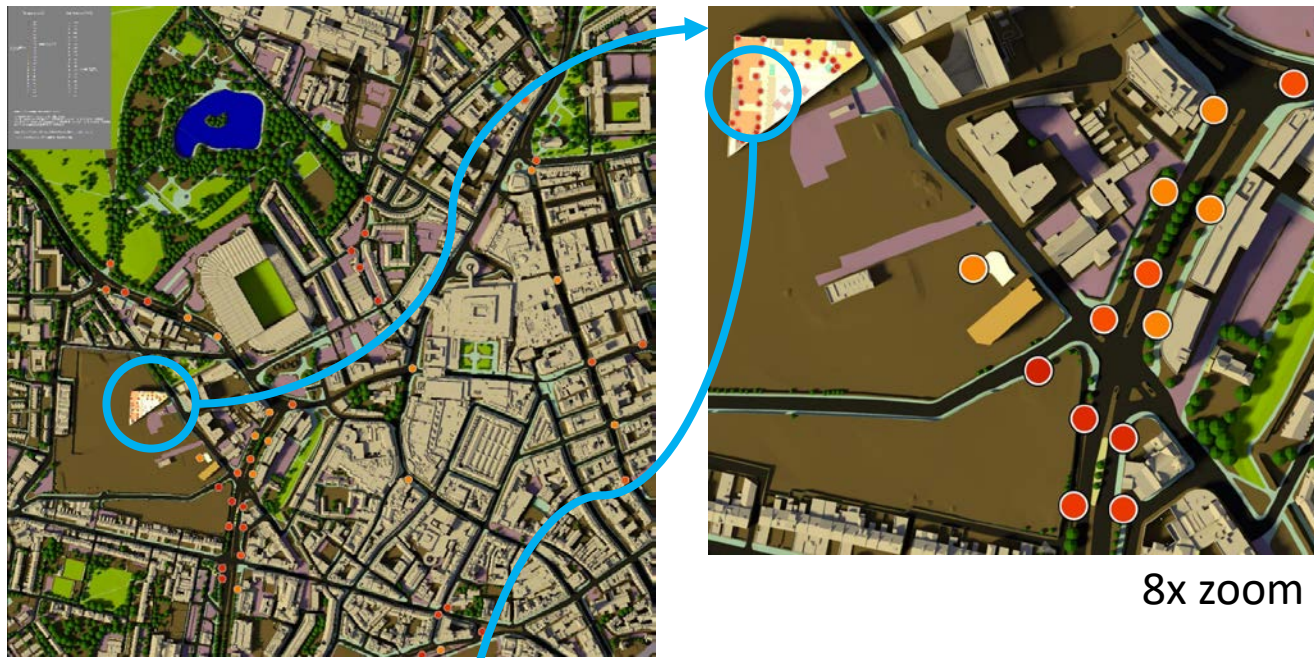
1048576 x 1048576 pixels

Zoomable 512x to full HD:

2.5 m x 1.25 m

2048x1024 pixels

1x1 pixel ~1.22x1.22 mm



#TeraScope

What needs to be computed?

krpano Hierarchy Level	4k*4k images	512*512 tiles
	65,793	5,592,405
12	65536	4194304
11	16384	1048576
10	4096	262144
9	1024	65536
8	256	16384
7	64	4096
6	16	1024
5	4	256
4	1	64
3		16
2		4
1		1

Render **65,793** $4096*4096$ images,
i.e. all pixels in levels 12, 8 and 4.

Then build a hierarchical image data
set of **5,592,405** $512*512$ pixel tiles.



#TeraScope

Live IoT data

Task Queue
Management
Azure Batch

65,793
4k Render Tasks



Parallel Rendering and
Image Splitting  

N up to 1024
GPU nodes

Parallel Rendering and
Image Splitting  



1Tbyte Azure Blob
storage per image

krpano
Tile Storage

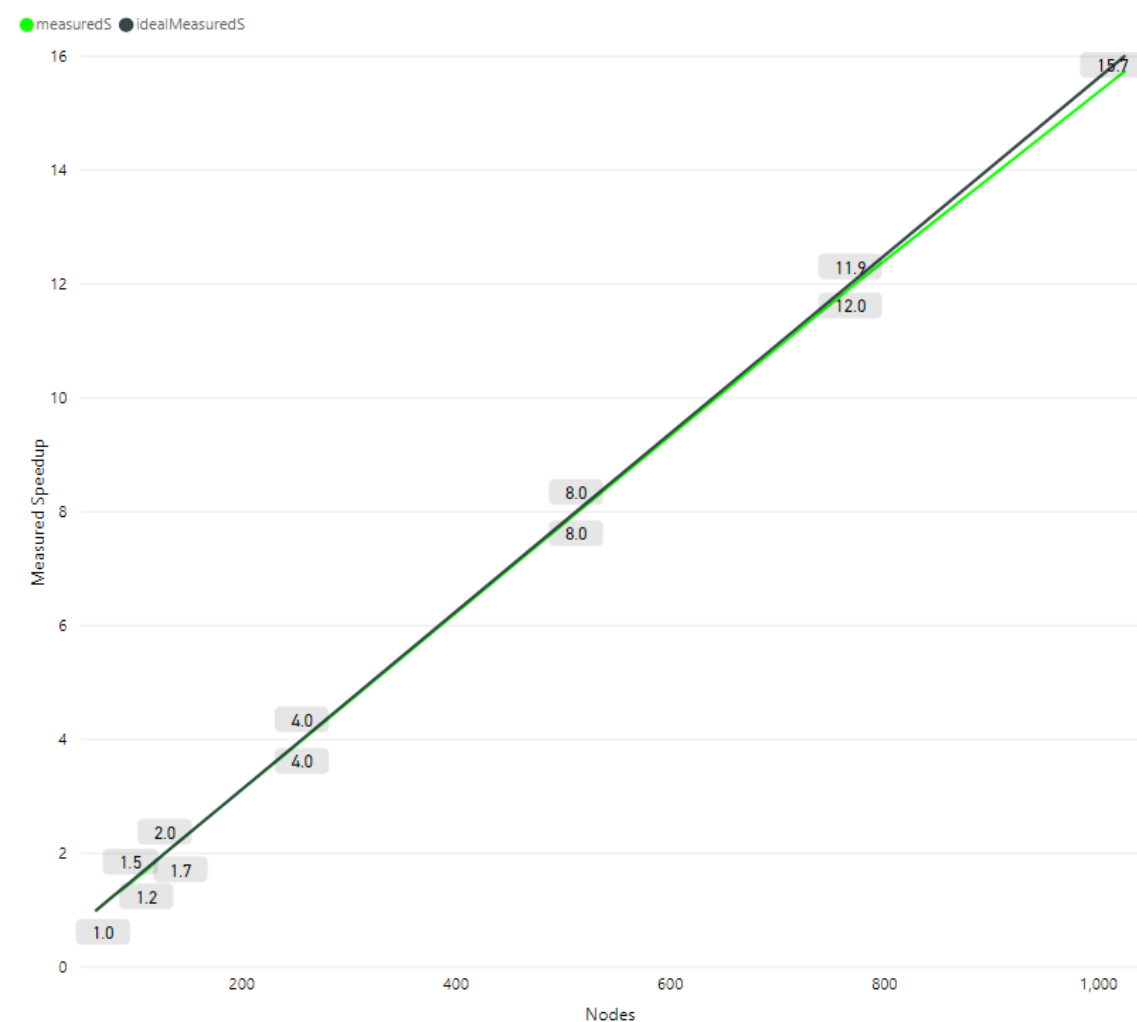
HTTP
Server

5,592,405
512x512 Image Tiles

Thin Client
Browsers



Peak **14 Pflop** visual supercomputer: 1024 NC6v3 (6 core + 1 Tesla V100 GPU) nodes plus 1 Tbyte of Azure blob storage.



Terapixel V100: measured speedup

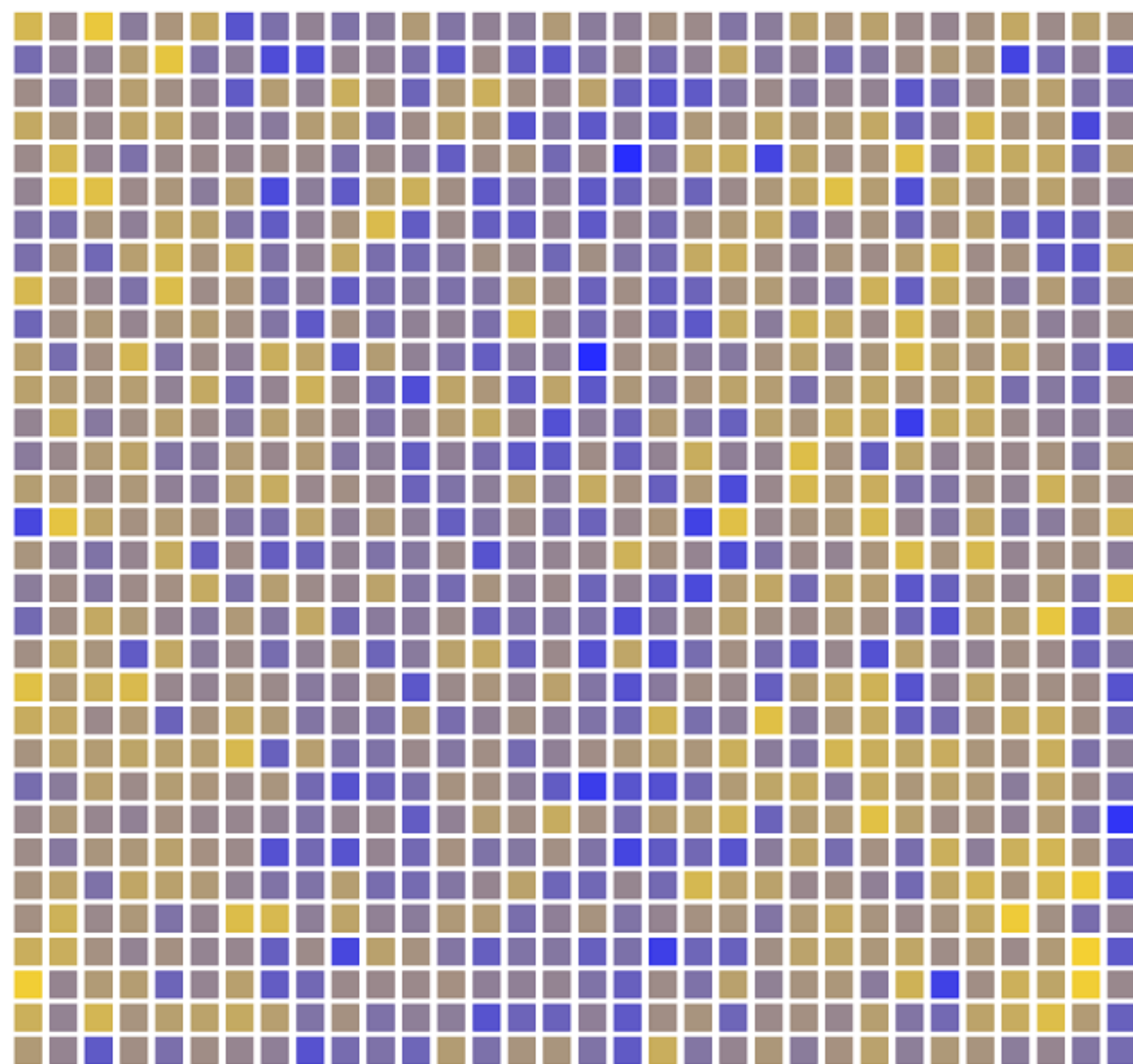
uniqHosts	measuredS	idealMeasuredS	measuredEf	totRunTme
1,024.00	15.7	16.0	0.98	2,901
768.00	11.9	12.0	0.99	3,824
512.00	8.0	8.0	1.00	5,726
256.00	4.0	4.0	0.99	11,469
128.00	2.0	2.0	1.00	22,778
112.00	1.7	1.8	0.98	26,551
96.00	1.5	1.5	0.98	30,900
80.00	1.2	1.3	1.00	36,567
64.00	1.0	1.0	1.00	45,632

Scaled compute from 1 to 1024 NC6v3 in East US region, peak of 14 Pflops.

Rendering time dropped from (estimated) 32 days on one GPU to 48 min. on 1024 GPU.

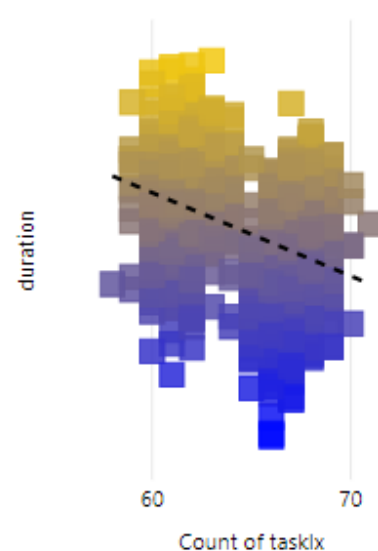
Cost: per run < £1.5K vs cost of the machine > £10M

#TeraScope : 1024 GPU node performance by total of GPU render time



Each tile represents a single V100 GPU.

- More GPU time on render tasks
- Less GPU time on render tasks



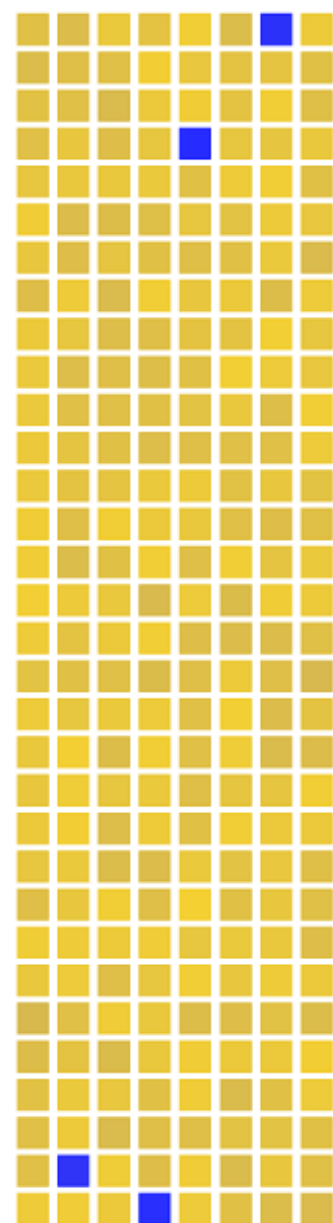
Overall we expect (and see) a trend that the more tasks a node computes the less GPU time is spent rendering.

High task counts result in higher overheads as image splitting and file store operations run more often.



Select ventile(s) of time to see GPU time per node

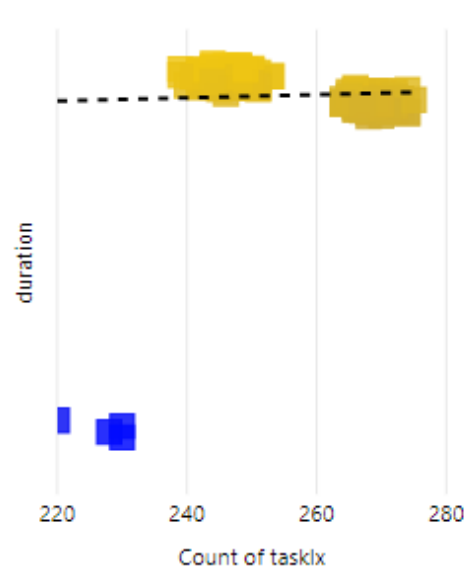


#TeraScope : 256 GPU node performance by total of GPU render time



Each tile represents a single V100 GPU.

-  More GPU time on render tasks
-  Less GPU time on render tasks



Overall we expect (and see) a trend that the more tasks a node computes the less GPU time is spent rendering.

High task counts result in higher overheads as image splitting and file store operations run more often.

Select ventile(s) of time to see GPU time per node

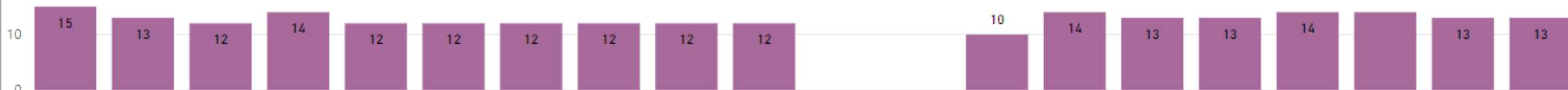


#TeraScope : 256 Tasks computed by GPU time and by count of tasks (outliers)

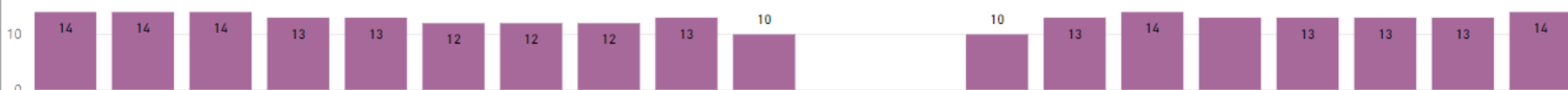
Number of tasks computed per ventile of run time



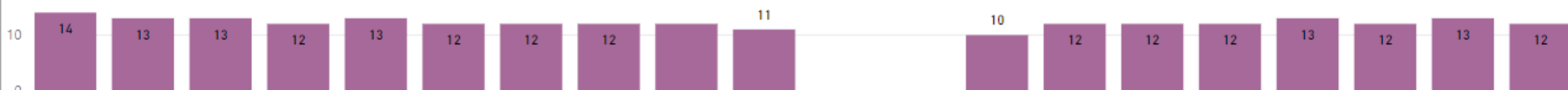
Number of tasks computed per ventile of run time



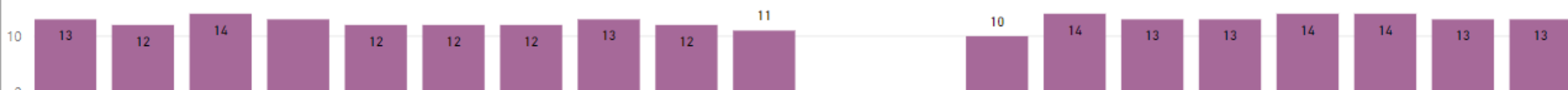
Number of tasks computed per ventile of run time



Number of tasks computed per ventile of run time



Number of tasks computed per ventile of run time



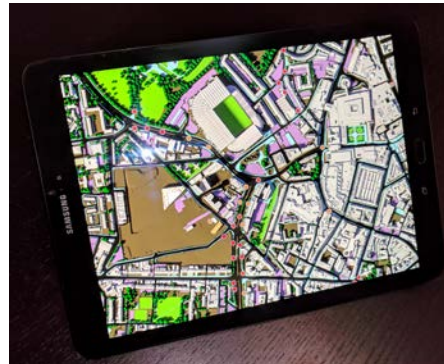
#TeraScope: Terapixel accessibility on any web enabled device



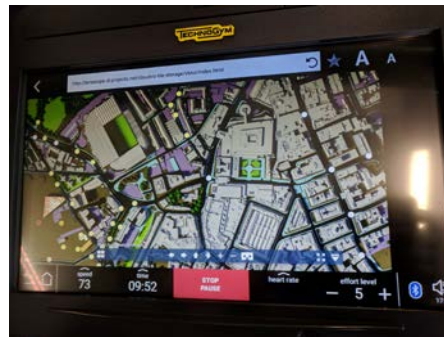
Curtin HIVE (Perth, WA)
180 deg 3m x 8m projection wall



Smart phone Pixel 2 XL



Samsung Galaxy Tab S3



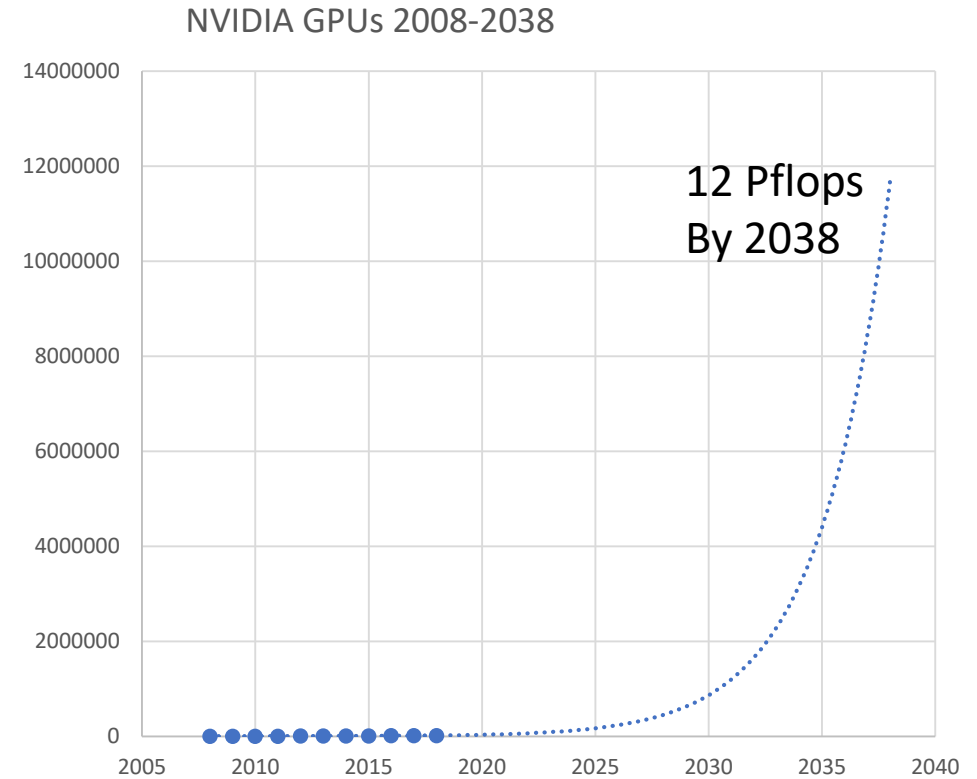
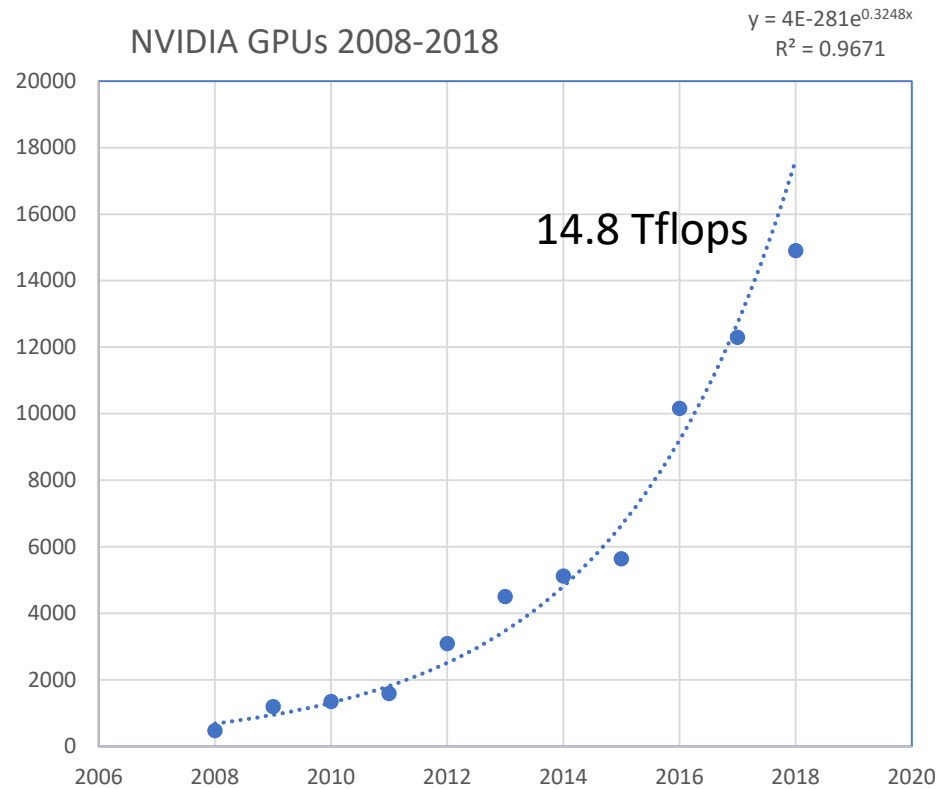
Gym cycling machine



Exeter Digital Humanities Lab
18M pixel LCD display wall

<http://terascope.di-projects.net/cloudviz-tile-storage-1024/vtour/index.html>

#TeraScope: How far into the future are we time travelling?



In the last ten years NVIDIA GPU performance has increased ~ 32x

Being able to use 1024x performance (14 PetaFlops) now provides us a platform that seems unlikely to appear on desktops until at least 2038.

#TeraScope Project outcomes

Visualization methods

panoramic images show potential for visualizing multiscale urban data.
accessible route to supercomputer visualizations on many low cost devices.

Visual supercomputing

deployed 14 PFlop cloud supercomputer, larger than any GPU HPC system in UK.
cost in total for one result (a scaling graph) ~£20,000 using > £10 million computer.
Azure improved (from K80 to V100) during the project, we immediately benefited.
provides access to performance approx. 20 years ahead of current desktop systems.

#TeraScope lets you zoom 512x into a trillion pixels on any web browser,
go to links.di-projects.net and interact with the TeraScope.

#TeraScope: automating data visualization

Visual entropy as a universal representation of uncertainty

Sara Fernstad (Newcastle, Computing)

Andrew Woods (HIVE, Curtin, Perth, WA)

Jenny Read (Newcastle, Neuroscience)

Arzu Coltekin (Geography, Zurich, Switzerland)

Darren Wilkinson (Newcastle, Maths and Statistics)

Kevin Wilson

Thank you

#TeraScope: Cloud Super-computing

Mark Turner (Newcastle, Computing)

Stephen Dowsland (Newcastle, Computing)

James Charleton (Northumbria, VNG)

Manu Anthony (Newcastle, Computing)

Phil James (Newcastle, Engineering)