NSE Big Data Challenge

Time	Activity
2.00pm	Introduction to NSE Big Data Challenge
2.10pm	What is Data Processing? - Sharing on the parameters collected and explanation of the processed data
2.30pm	Demonstration on ModStore
3.00pm	Sharing on the Supercomputing Facilities
3.10pm	Hands-on Session for Students
4.00pm	End of Workshop

National Research Foundation
Prime Minister's Office
Singapore

Institute of High Performance Computing
A*STAR

Singapore University of Technology and Design

Science Centre Singapore

NSCC Singapore

Background and Objectives

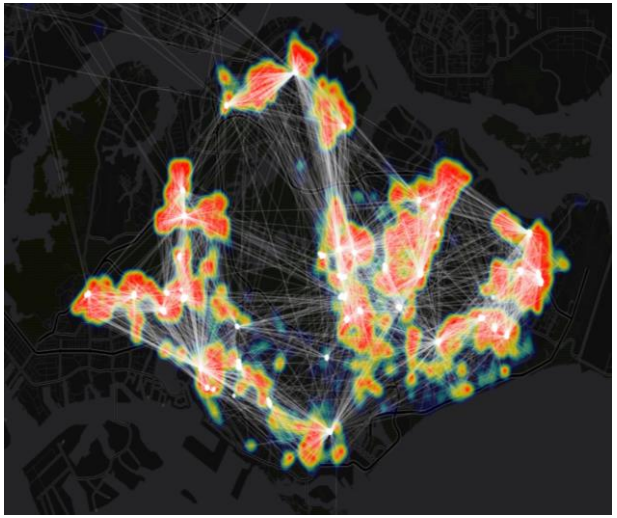
- A **nation-wide project** launched by President Tony Tan in Jan 2015
- First National-scale deployment of **IoT devices** designed for ease of use
- Involved **176 schools and more than 90,000 students,** and in 2 years

NSE Big Data Challenge

Objectives:

To allow students to learn about big data analytics through the use of the NSE data

- Processing and filtering of big data
- Use of big data tools
- Draw meaningful insights from big data
- Presentation of analyses in easy-to-understand ways



From Data to Decisions

Timeline

S/N	Date	Date
1	Half-day Preparatory Workshop	17, 18, 19 Oct
2	Submission of Entries	9 Dec 2016
3	NSE Big Data Challenge Finale - Exhibition, Prize Ceremony	3rd week of Jan 2017

Prizes

1st prize:

- Up to **$300** worth of gifts for each member

+ cash contribution to school student fund

2nd prize:

- Up to **$200** worth of gifts for each member

+ cash contribution to school's student fund

3rd prize:

- Up to **$100** worth of gifts for each member

+ cash contribution to school's student fund

Consolation prizes

- Up to **$50** worth of gifts for each member

+ cash contribution to school's student fund

Materials to be submitted

1. ***Written Report*:***
   a) *Innovation (25%)*
      - How creative is the use of the data
   b) *Technical Accuracy (25%)*
      - How well is the data processed
   c) *Impact (25%)*
      - How much social, environment, economic value does it create
2. ***Presentation of Analyses***
   a) *Using maps, slides, video, etc. (25%)*

*6-page report, font size 12, including annexes

What Data is Available?

Sensor	Range	Accuracy	Units	Poll Freq (Hz)
Accelerometer	±2g~±16g	-	m/s^2	100 (for 1 sec)
Gyroscope	±250 to ±2000	-	deg/sec	100 (for 1 sec)
Magnetometer	±4800uT	-	uT	100 (for 1 sec)
Light Intensity	0.165 to 100k	-	lux	0.1
Sound pressure	30 to 130	SNR: 63	dB	0.1
Relative Humidity	0-100	+/- 3	%	0.1
Amb.Temperature	-10 to +85	+/- 0.3 @ 25°C	°C	0.1
Pressure	300 to 1100 hPa	+/- 0.12 hPa	hPa	0.1
IR Temp	-40 to 125	+/- 3	°C	0.1
Buzzer	-	-	-	-
RGB LED	-	-	-	-
Wi-Fi Radio	-	-	-	-

How was Data Processed?

Variable	Explanation
aircon_co2	CO2 emissions from aircon
aircon_energy	Energy consumption of aircon
poi_lat	Point of interest (POI) latitude
poi_lon	Point of interest (POI) longitude
stairs _climbed	Number of stairs climbed
travel_co2	CO2 emissions from the transport mode
outdoor_time	Time spent outdoor
am_travel_mode	Transport mode in the morning
pm_travel_mode	Transport mode in the afternoon

Air-con usage

Identified by a temperature threshold + rapid drops/rises in humidity to mark the start & stop times

PoI

identified if a number of points cluster in a particular area, e.g. 4 points in a 2min span around the same location. School/home PoIs guessed based on time of day, shopping centres, etc. based on any such PoI found in vicinity on Google Maps.

Stairs

Identified by pressure differentials (just like the air-con usage is identified by humidity differentials).

Travel modes

Walking trips identified based on speed threshold (e.g. 1m/s), other transport modes also based on speed, and accelerometer patterns. Public transport trips based on number of points along a public transport route using Google Maps.

Outdoors time

Differentiated by light intensity; bright = outdoors, dark = indoors.

What is ModStore?

Institute of High Performance Computing
A*STAR

Data science platform to explore, visualize and find insight from data. **All on a browser!**

- No need to download data – access your work anywhere!
- No need to write formula
- Included a wide range of statistical and visualization tools, e.g. histogram, boxplot, t-test, correlation, heat map
- Just drag and drop the appropriate tools to build your own model

Process of data investigation

Institute of High Performance Computing
A*STAR

Iterative process:

1. What data is available?
2. What are possible problems my team can pose using this data?
3. Explore data
4. Present findings and solutions

What data is available in ModStore?

Institute of High Performance Computing
A*STAR

Your own school's **raw** data

- Each time a device uploads raw sensor data, it creates one row in the dataset
- E.g. temperature, humidity, noise

Your own school's **processed** data

- Each row in the dataset represents the processed data for each experiment day
- E.g. transport mode, distance and duration in the morning and afternoon

Your own school's **happy button** data (2016 only)

- Each time the happy button is pressed, it creates one row in the dataset
- 1 = happy, 2 = happier

Example: raw data

Institute of High Performance Computing
A*STAR

#	id	date	time	humidity	light	mode	noise	pressure	steps
21	507202	2016-07-11	08:34:59	64.5	0	1	55	100998	96040
22	507202	2016-07-11	08:35:22	64.5	0	1	53	100998	96040
23	507202	2016-07-11	08:35:45	64.5	0	1	54	101003	96040
24	507202	2016-07-11	08:36:08	64.5	0	1	54	101001	96040
25	507202	2016-07-11	08:36:31	64.5	0	1	53	101003	96040
26	507202	2016-07-11	08:36:54	64.5	0	1	51	101000	96040

Example: processed data

Institute of High Performance Computing
A*STAR

#	id	date	aircon_co2	aircon_energy	am_travel_distance	am_travel_mode	am_travel_duration	pm_travel_distance
1	708784	2016-07-11	3028.5	7.0105			"[0.4789, 2.8662, 0.1074, 15.233, 2.0704]"	Bus
2	708784	2016-07-12	5902.5	13.663	"[0.6201, 13.692, 0.3266, 3.6756, 0.1702]"	Car	"[585, 1275, 348, 948, 97]"	"[0.2799, 4.2103, 0.3277, 15.577]"
3	708784	2016-07-13	6428.1	14.88	"[0.5827, 13.8, 0.3143, 3.8823, 1.3147]"	Bus	"[427, 1376, 276, 1309, 574]"	"[0.1666, 15.992, 0.0748, 0.4542]"
4	708784	2016-07-14	6531.1	15.118	"[0.5269, 14.427, 0.2074, 3.9057, 1.1816]"	Bus	"[539, 1390, 313, 1161, 885]"	"[0.198, 3.5975, 0.3222, 14.908, 0.1151]"
5	708584	2016-07-11	1741.5	4.0314			"[0.0915, 2.7039]"	"[116, 312]"

Example: happy button data

Institute of High Performance Computing
A*STAR

#	id	date	time	button	info
1	708784	2016-07-11	10:29:38	1	DURATION:2
2	708784	2016-07-11	19:55:42	2	DURATION:1
3	708784	2016-07-13	08:33:28	2	DURATION:1
4	708784	2016-07-13	15:32:08	2	DURATION:362
5	708784	2016-07-13	15:32:30	2	DURATION:478
6	708784	2016-07-13	15:33:29	2	DURATION:1
7	708784	2016-07-14	09:56:56	1	DURATION:755
8	708584	2016-07-11	13:58:16	2	DURATION:1

Problem posing

- What are possible problems my team can pose using this data? (must be relevant to criteria)

- Audience for whom the data is interpreted (e.g. school management, LTA, NEA, society at large)

Examples: problem posing

Institute of High Performance Computing
A*STAR

- How much time does a private car save me on the road compared to public transport? What's the amount of CO_2 reduction if the whole school switches to public transportation?

- How much outdoor time do students in my school spend? Are there other factors (e.g. haze, Pokemon Go) contributing to this activity level?

- Define a new measure e.g. “resilience quotient”. Choose suitable variables to support definition, e.g. mode of transport, air-con usage from humidity and temperature data, outdoor time, step count, stairs-climbed and use the data to draw some conclusions

Exploring the data

- Measures of central tendency, spread
- Clustering the data (binning)
- Ways to represent the data (e.g. types of graphs)

INFERENTIAL STATISTICS
- PolyFit
- CorrelationCoefficient
- T-test_ind
- T-test_paired
- Z-test
DESCRIPTIVE STATISTICS
- Quartile-range
- Mean-median-mode
- Variance

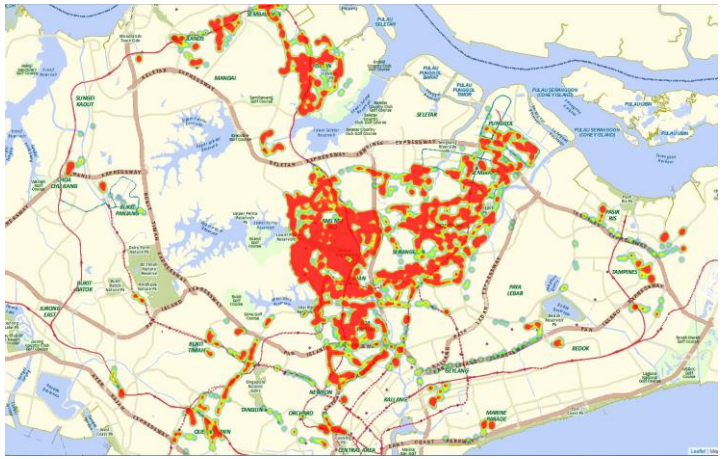
DIAGRAMS
- Barchart
- BoxPlot
- Cumulative
- Histogram
- DotDiagram
- Pictogram
- StemLeafPlot
- HeatMap
- Map
- HeatMapAnimation
- StackBar
- MultipleLines
- Linechart
- Piechart
ScatterPlot

Presenting findings and solutions

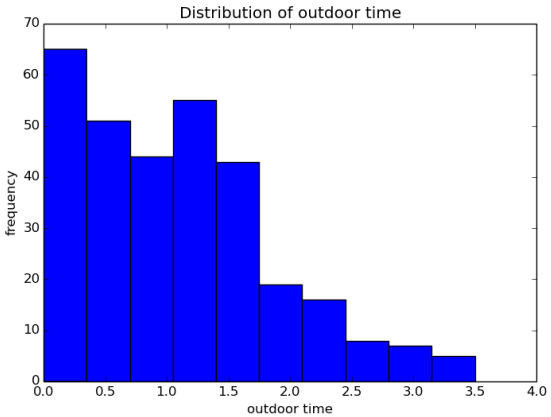
Institute of High Performance Computing
A*STAR

Videos, slides, maps…

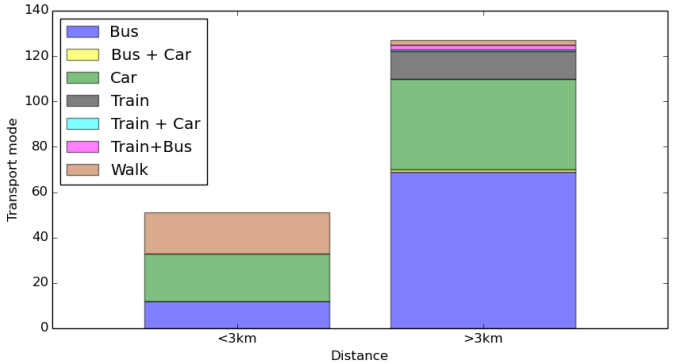
Heat map



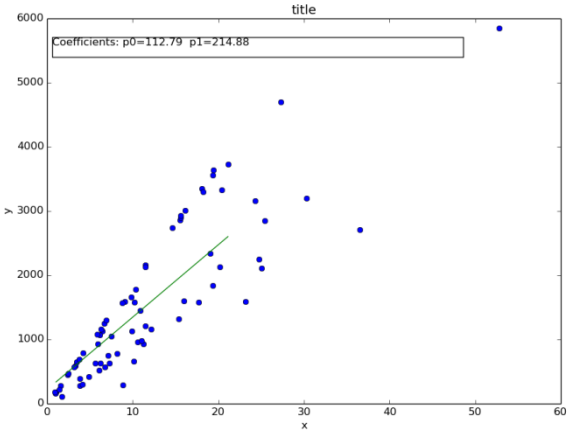
Histogram



Stacked Bar Chart



Curve fitting



ModStore step-by-step demo

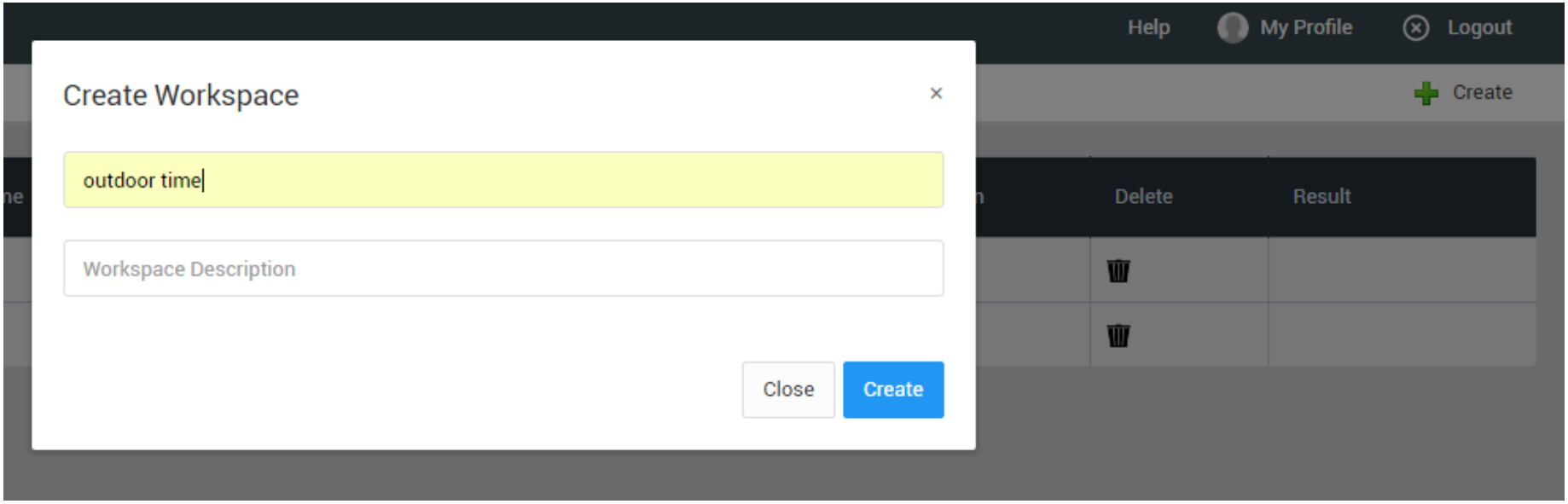
Institute of High Performance Computing
A*STAR

1. How much outdoor time do students in my school spend?

2. Estimate how much sleep students in my school are getting, and compare this to the recommended hours for teenagers

Demo 1: outdoor time

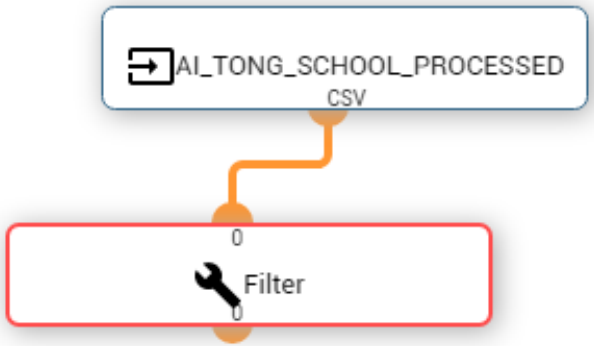
Institute of High Performance Computing
A*STAR

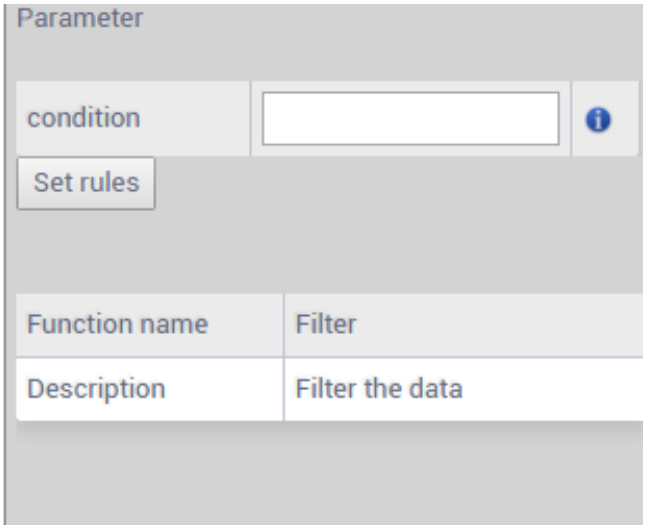


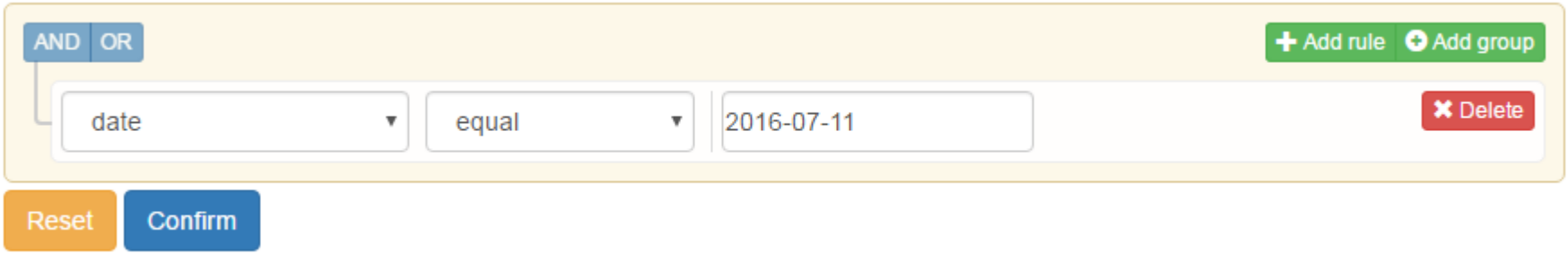
File name	AI_TONG_SCHOOL_PROCESSED
Extension	csv
Description	undefined
File size	238294
uri Path	undefined
Upload date	undefined

Preview Data | Download Data

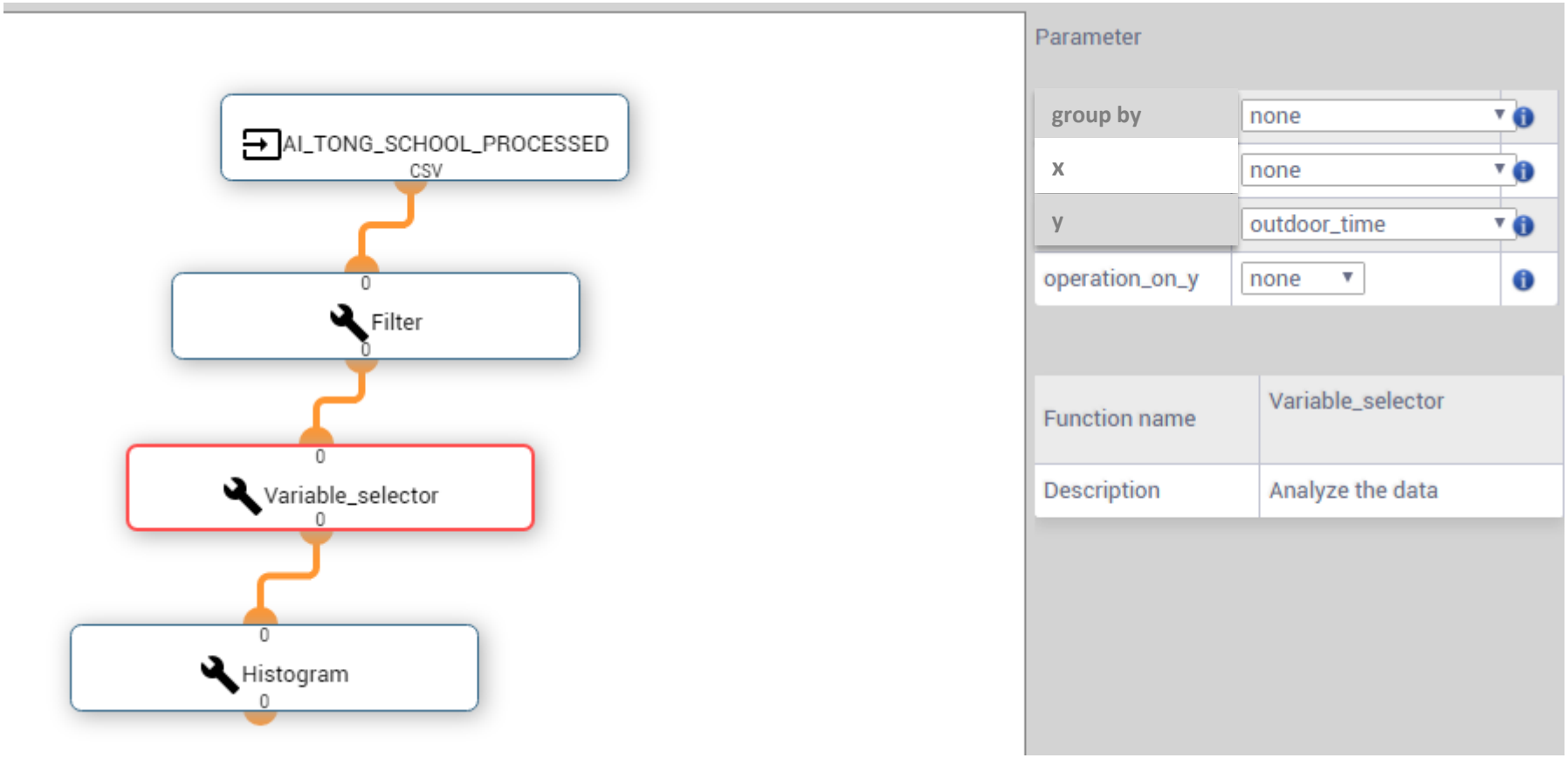
Demo 1: outdoor time



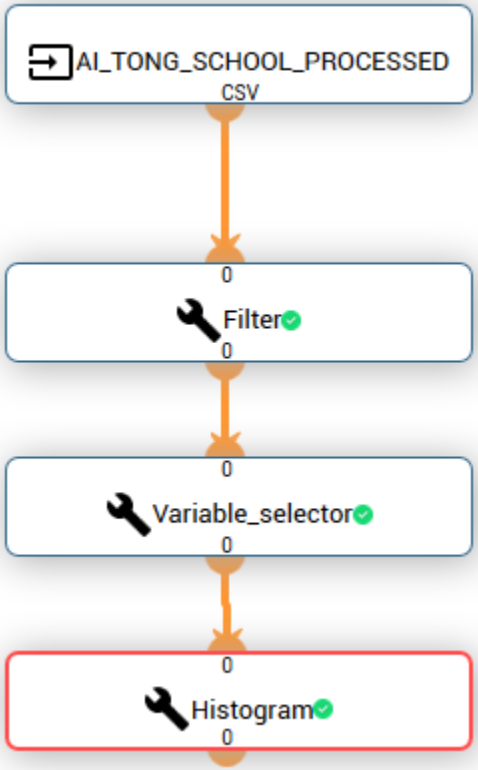


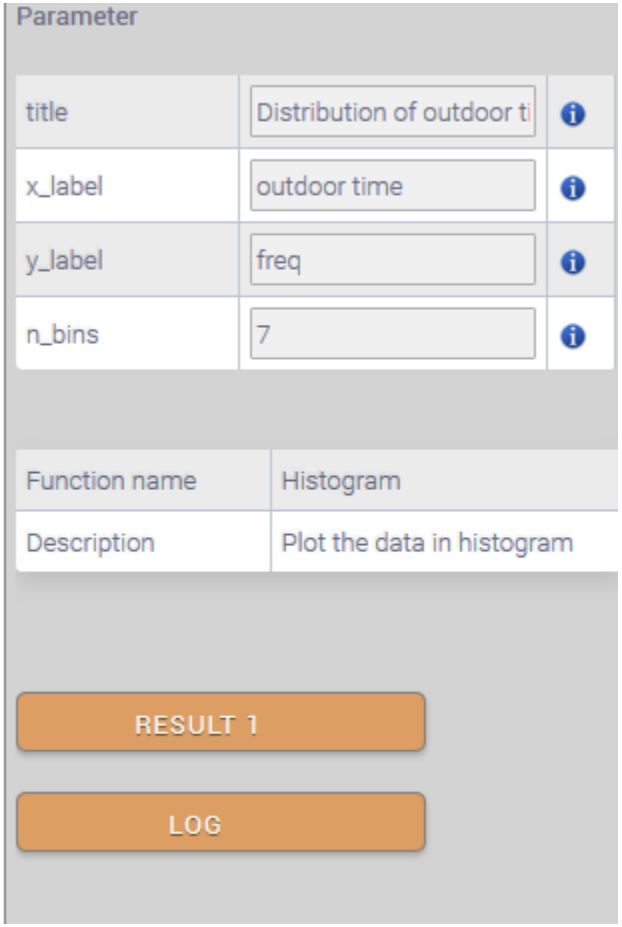


Demo 1: outdoor time

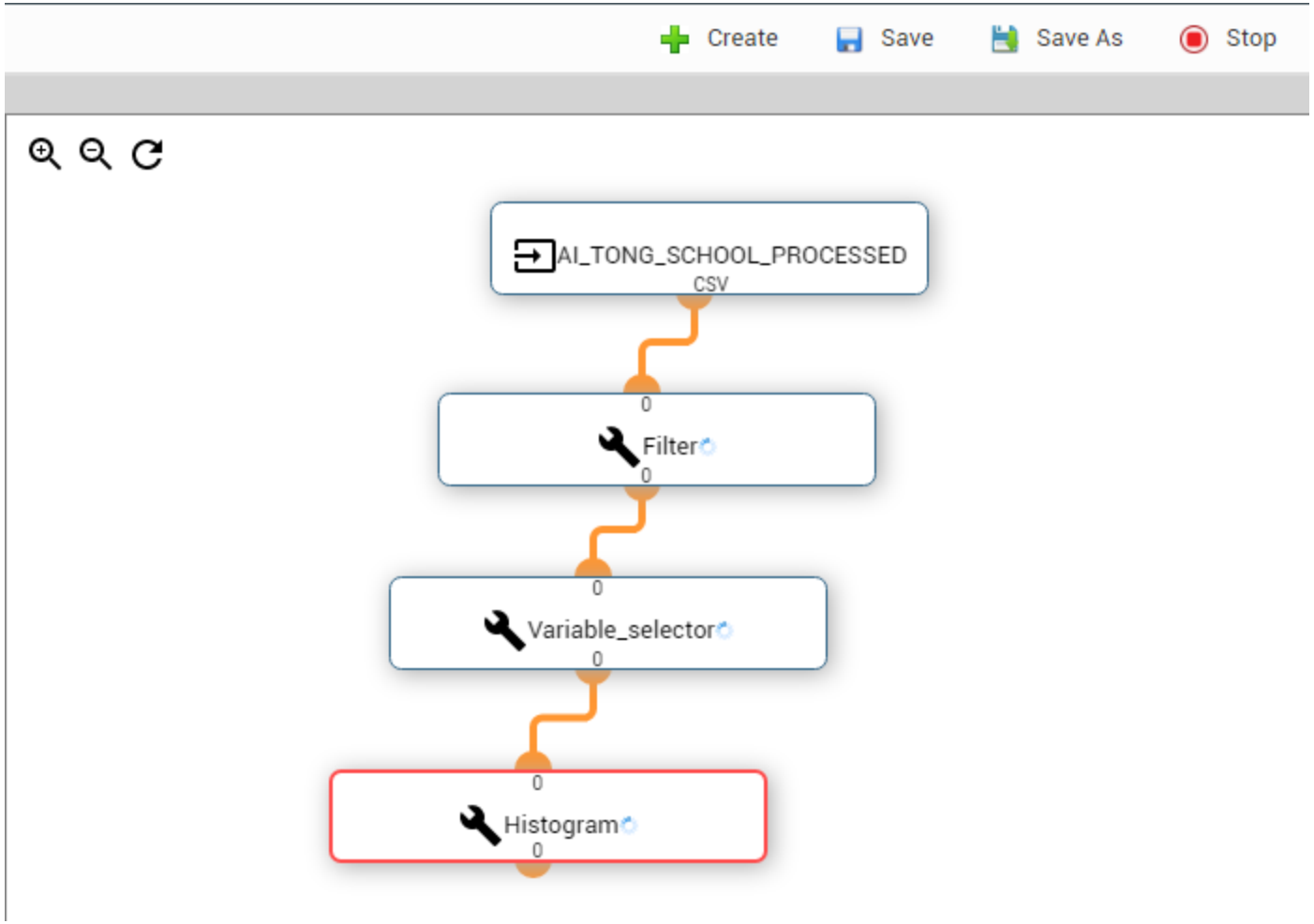


Demo 1: outdoor time

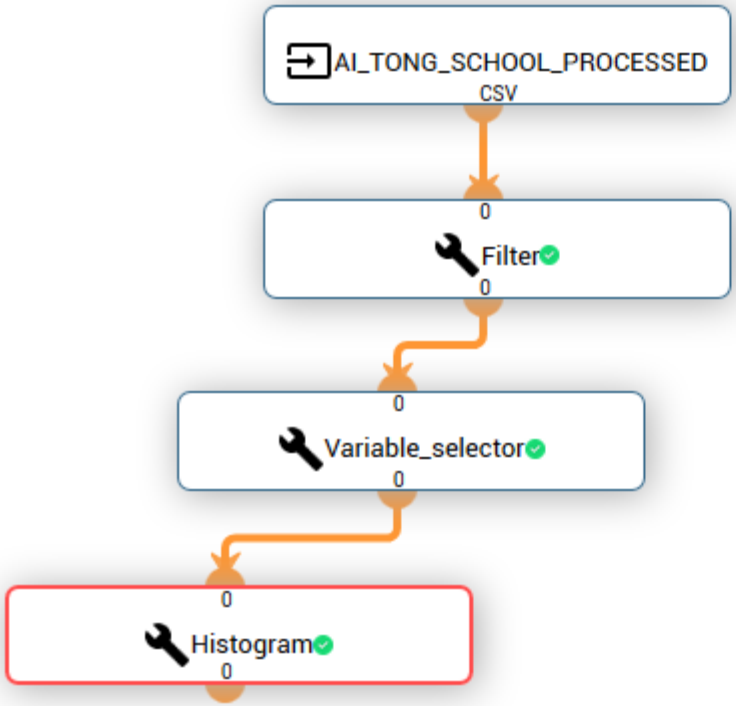


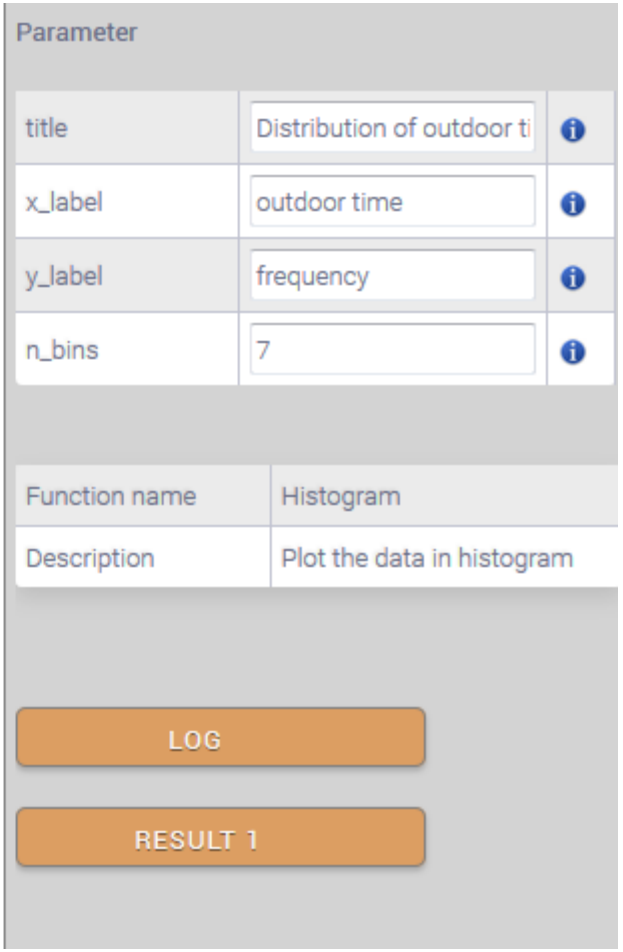


Demo 1: outdoor time

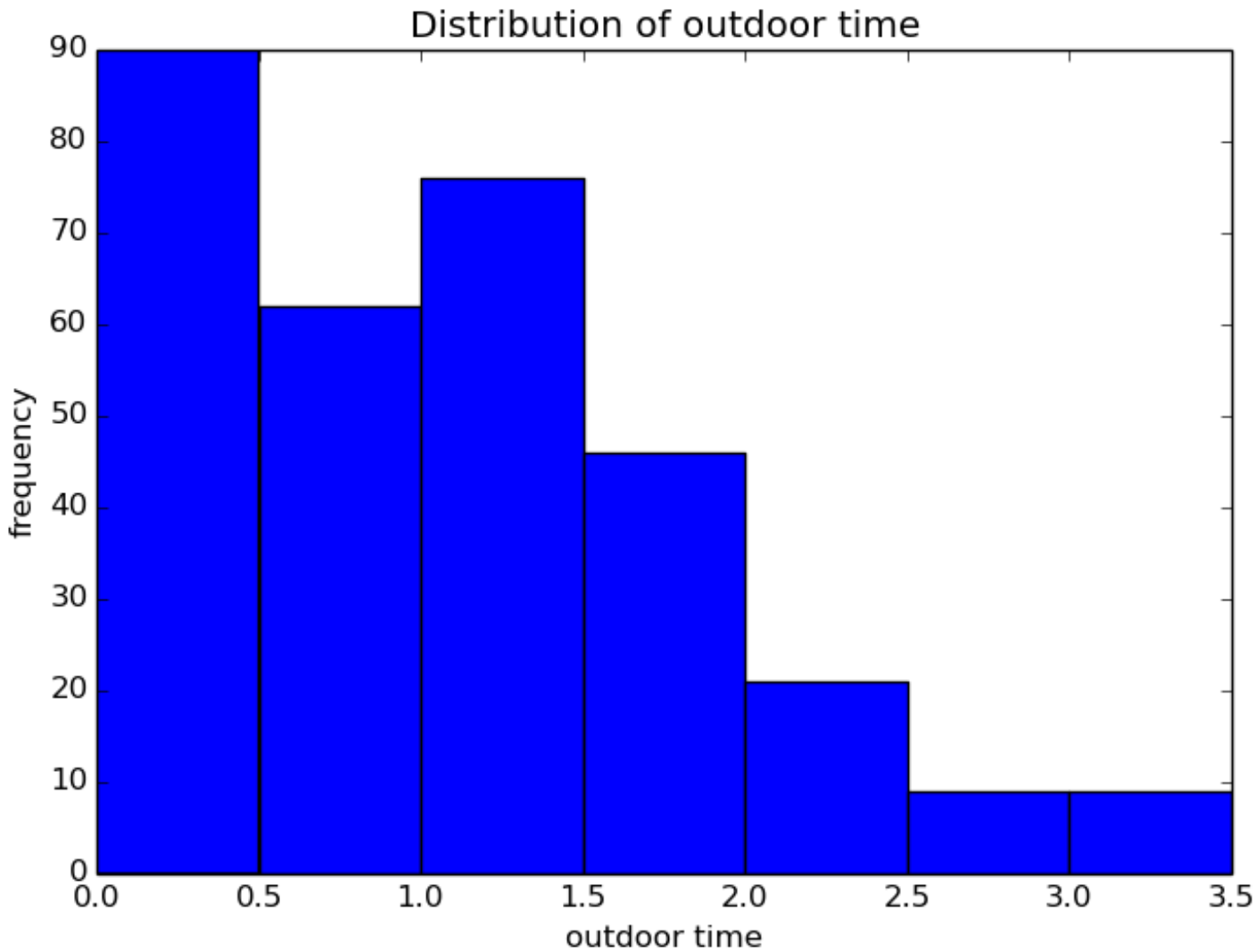


Demo 1: outdoor time





Demo 1: outdoor time

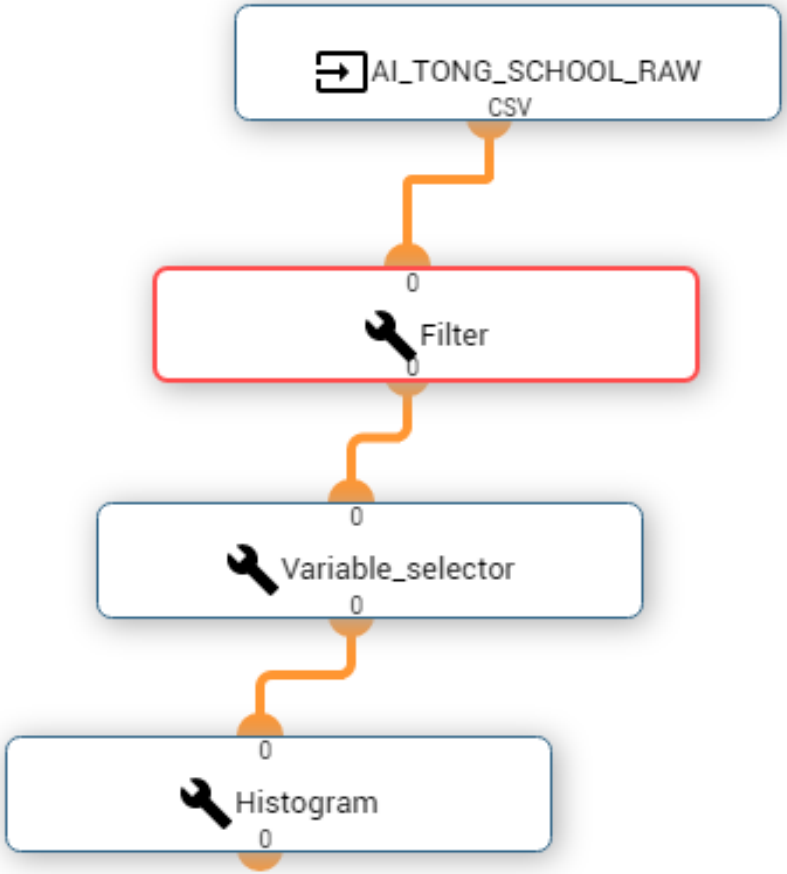


Demo 1: outdoor time

Institute of High Performance Computing
A*STAR

Display	Workspace name	Status	Date created	Last modified	Open	Delete	Result
	Outdoor time	Completed	14/10/2016 15:47:04	14/10/2016 15:58:36			
	12 July	Completed	14/10/2016 15:56:14	14/10/2016 15:58:36			
	11 July	Completed	14/10/2016 15:49:04	14/10/2016 15:50:45			

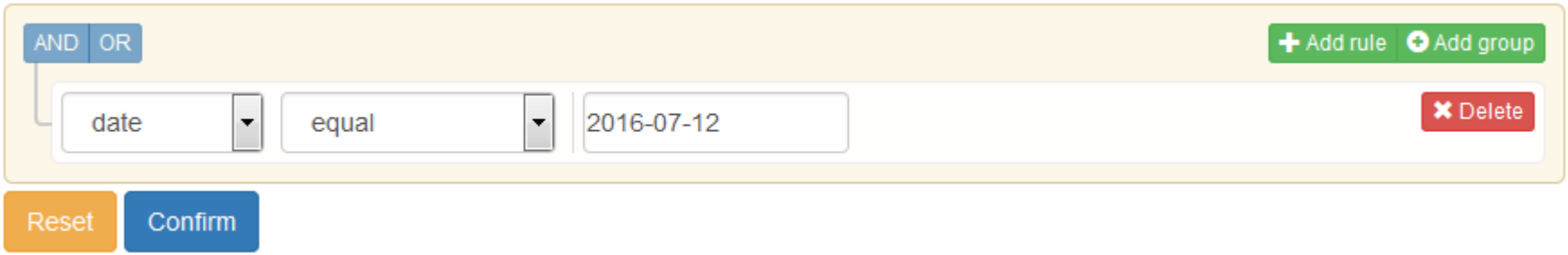
Demo 2: hours of sleep



Demo 2: hours of sleep

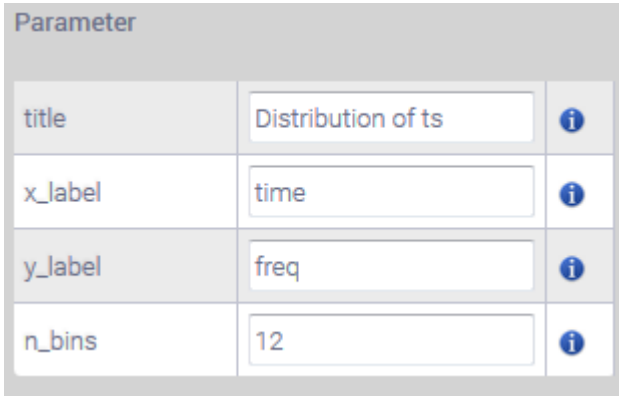
Institute of High Performance Computing
A*STAR

Filter



Variable_selector

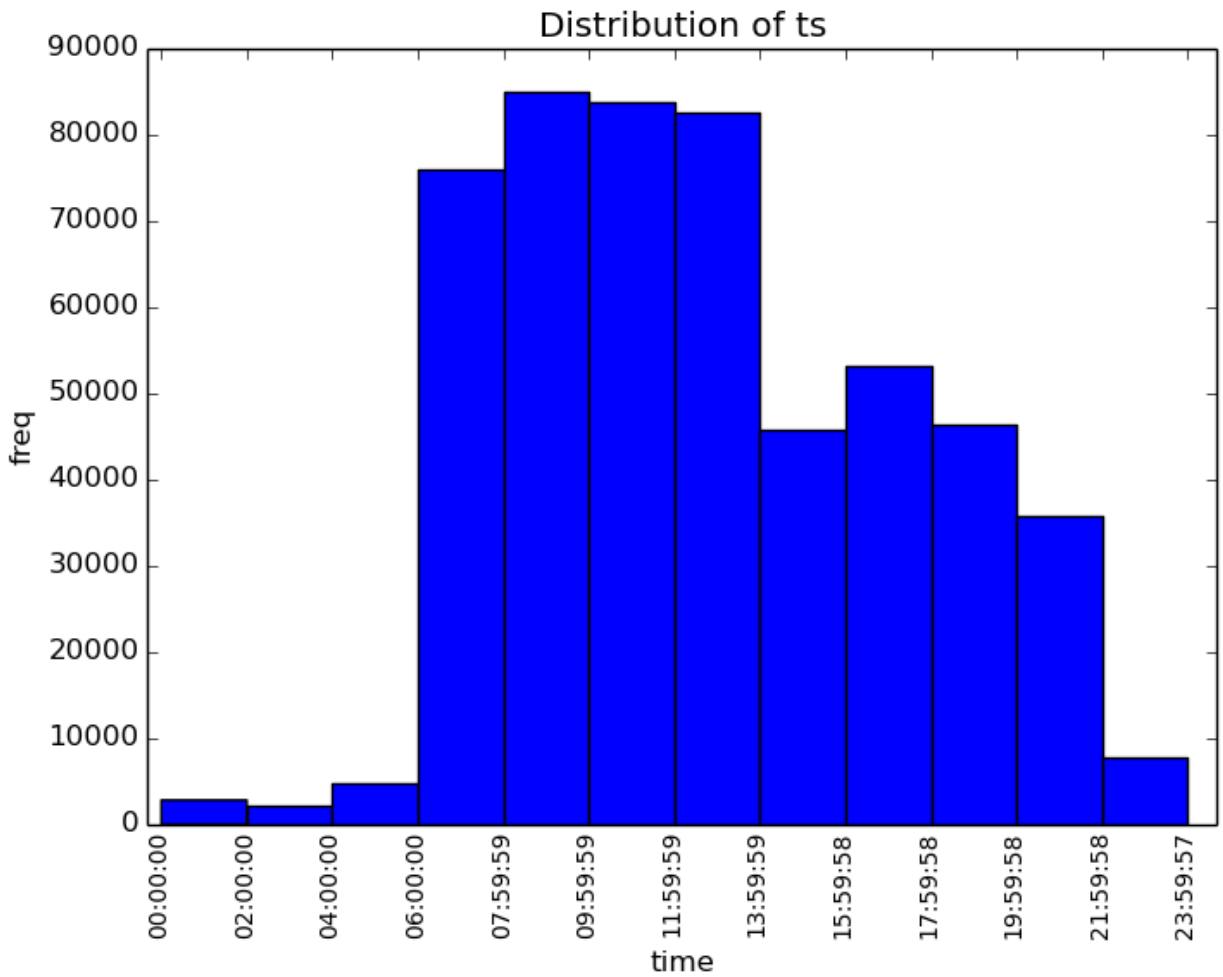
Histogram



Demo 2: hours of sleep

Institute of High Performance Computing
A*STAR

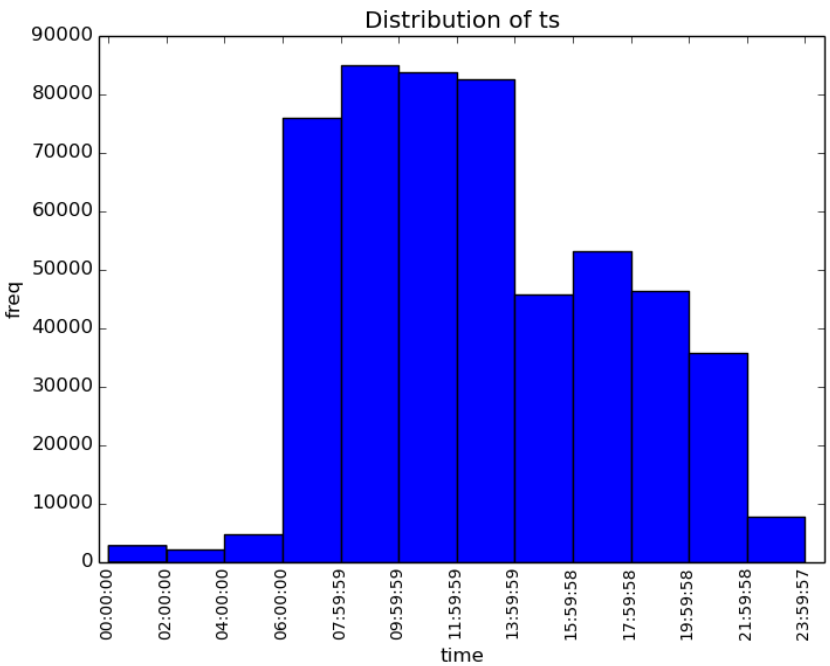
Total



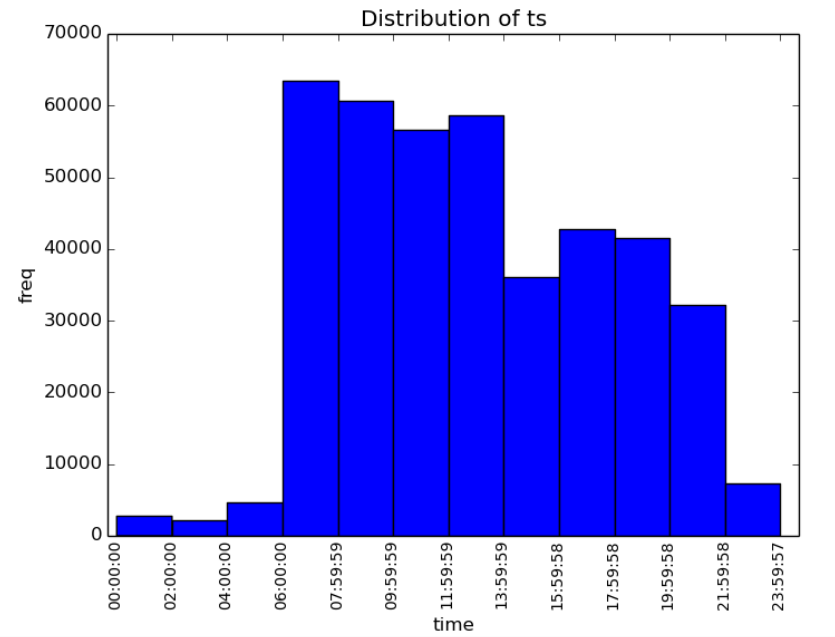
Demo 2: hours of sleep

Institute of High Performance Computing
A*STAR

Total



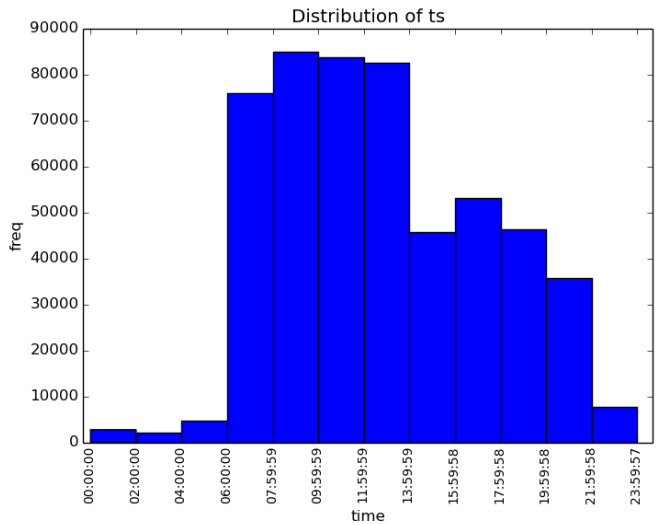
When light = 0



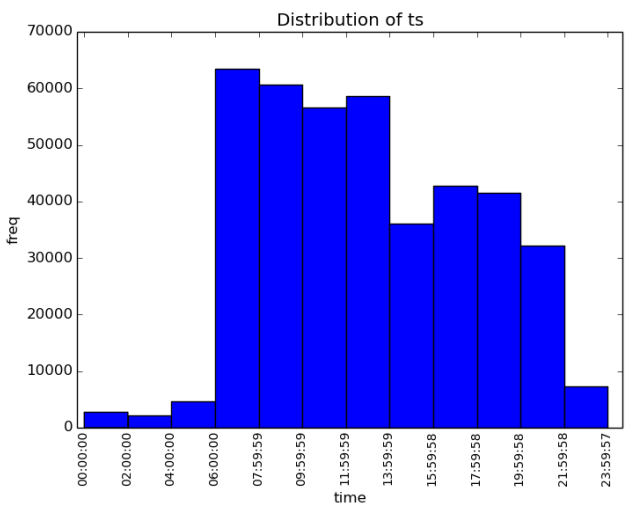
Demo 2: hours of sleep

Institute of High Performance Computing
A*STAR

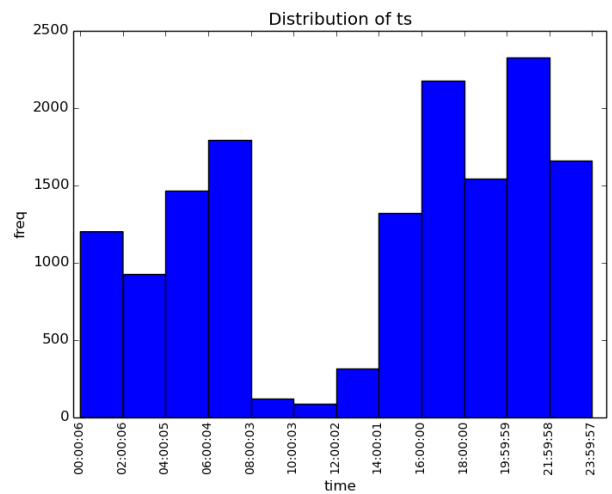
Total



When light = 0



When noise < 40

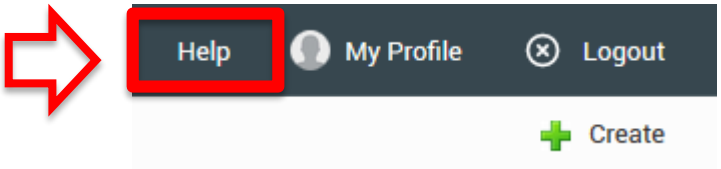


Important info: account

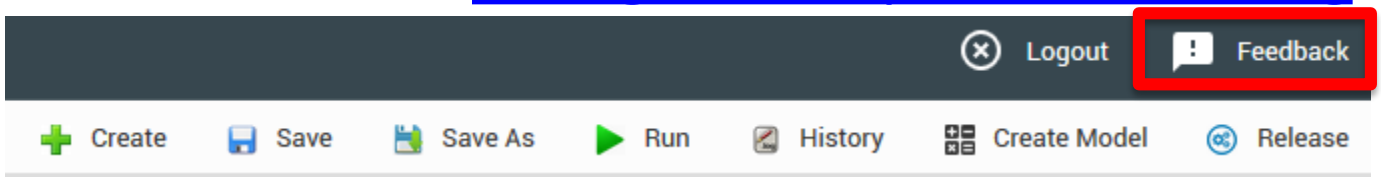
- Your ModStore username and password have been emailed to you before the workshop
- For those who did not provide an email address, your username and password will be emailed to one of your team members
- Login to http://modstore.org/nse/ to change your password
- Do not share account with your team members, as it could be confusing and problematic to save and run your model if different people work on it at the same time

Important info: help

- Before you start, watch tutorial and read user guide on "Help" page

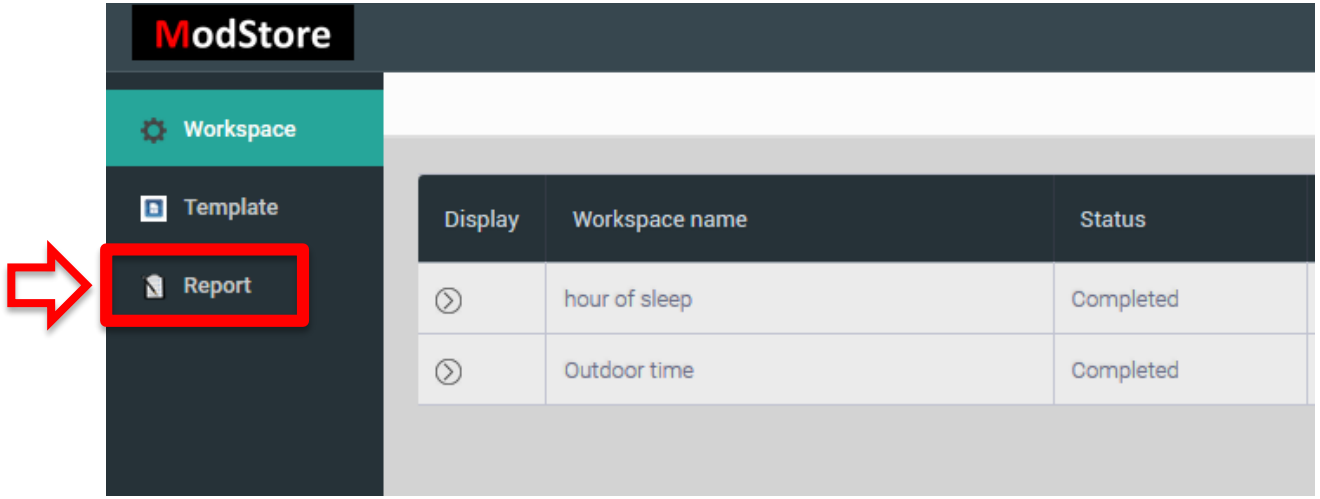


- When a problem occurs during use, first check Frequently Asked Questions (FAQ) section for possible solutions
- If you are unable to troubleshoot the problem, we are here to help. Submit the online feedback form, describe the problem in as much details as possible. We will try our best to respond to you in 2 working days.
- You can also email at zhangww@ihpc.a-star.edu.sg



Important info: submission

- Submit your final report and presentation via ModStore. Put all the materials in one zip file and upload.
- Only one member from the team needs to submit using his/her own account
- In case of multiple submissions within the same team, only the latest submission before the deadline will be used



National Supercomputing Centre (NSCC) Singapore

Presented by:
Ong Guan Sin
Head, New Services

Leong Wai Meng
Deputy Director (Business Development)

17- 19 Oct 2016

How do we support the NSE?

Urban Mobility and Environment Analysis

>90,000 students

200+ students accessing ModStore

> 250K CPU Core Hours

176 schools

17 Oct 2016 – 8 Dec 2016

> 150GB of Data

>300,000 km travelled

400 Million + lines of data

Who are we???

Our Stakeholders

Agency for Science, Technology and Research
SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY

NUS National University of Singapore

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

SG NSCC SINGAPORE

National Petascale Facility

The National Supercomputing Centre Singapore is a national Petascale facility established to support high performance science and engineering computing needs for academic, research and industry communities in Singapore

Introduction: Vision & Objectives

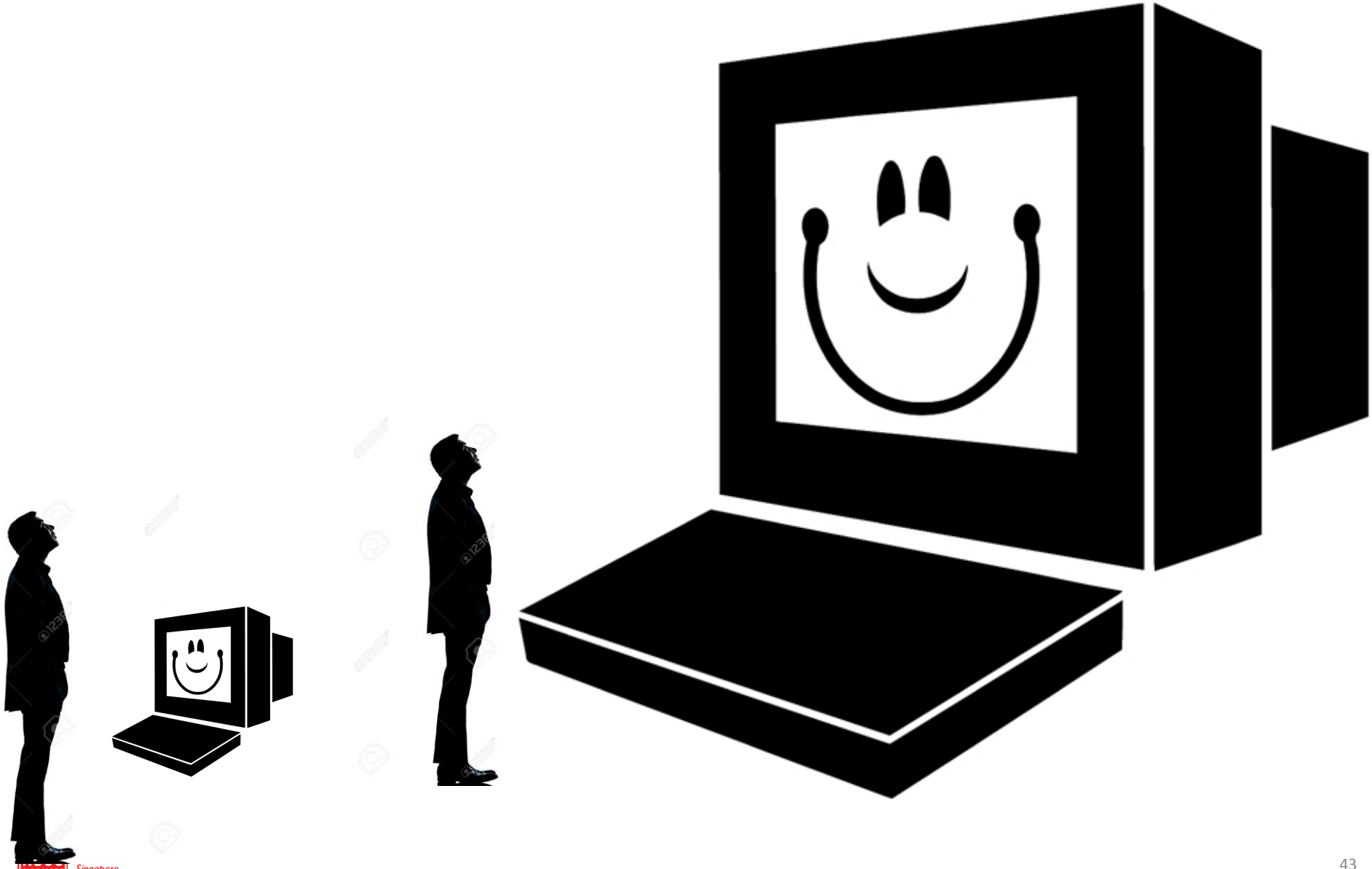
Vision of NSCC

"Democratising Access to Supercomputing"

Objectives of NSCC

1. ***Support National R&D Initiatives***
2. ***Attract Industrial Research Collaborations***
3. ***Enhance Singapore's Research Capabilities***

What is a supercomputer?



HPC Hardware

~1 PFLOP System

- ***> 30,912 CPU cores***
- ***1,288 nodes*** *(dual socket, 12 cores/CPU E5-2690v3)*
- ***128 GB*** *DDR4 RAM/ node*
- ***10 Large memory nodes*** *(1x6TB, 4x2TB, 6x1TB)*

FUJITSU

13PB Storage

- ***HSM Tiered, 3 Tiers***
- ***I/O 500 GB/s flash*** *burst buffer*
- ***10x Infinite Memory Engines*** *(IME)*

EDR Interconnect

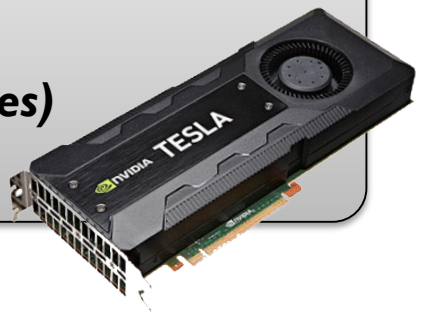
- ***EDR (100Gbps) Fat Tree within cluster***
- ***InfiniBand connection*** *to* ***remote login nodes*** *at stakeholder campuses (NUS/NTU/GIS) at 40/80/500 Gbps throughput*

Mellanox TECHNOLOGIES

HPC Hardware – Accelerator Nodes

Accelerator nodes

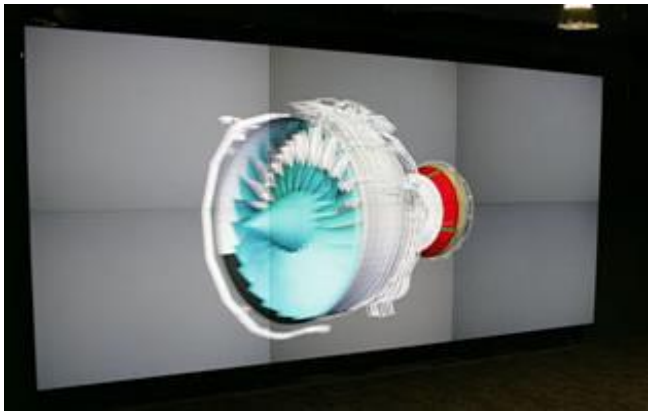
- ***128 nodes*** *with NVIDIA GPUs*
- ***NVIDIA Tesla K40 (2,880 cores)***
- ***368,640 total GPU cores***



GPU technology used for image classification, video analytics, speech recognition, and natural language processing.

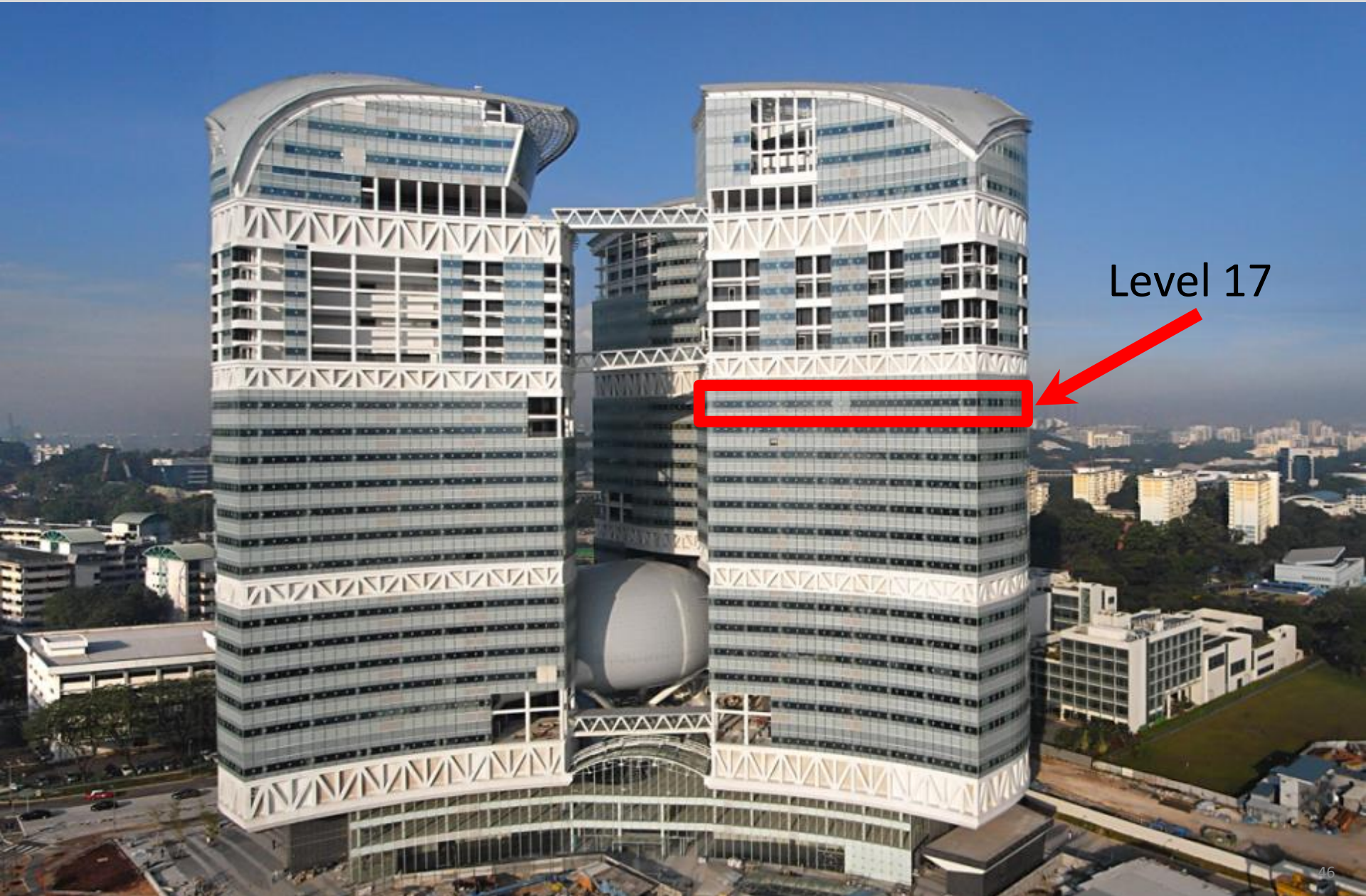
Visualization nodes

- ***2 nodes*** *Fujitsu Celsius R940 graphic workstations*
- *Each with* ***2 x NVIDIA Quadro K4200***
- *NVIDIA Quadro Sync support*



*[Image courtesy of A*CRC]*

NSCC Data Centre @ Fusionopolis



NSCC Data Centre

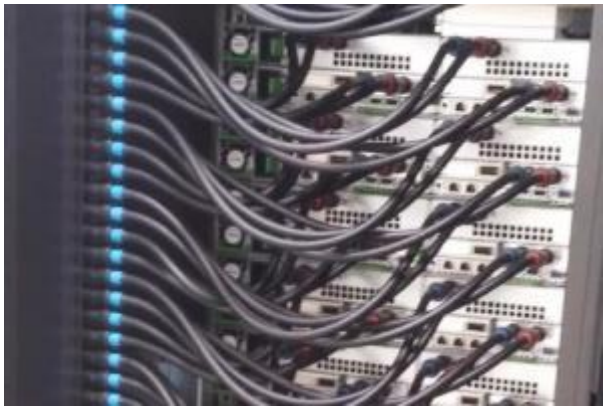
NSCC Data Centre – Cooling System

Combination of 3 cooling systems to achieve max. efficiency

Air Cooling:
Computer Room Air Handler (CRAH) units

Chilled water Cooling:
Rear door heat exchangers

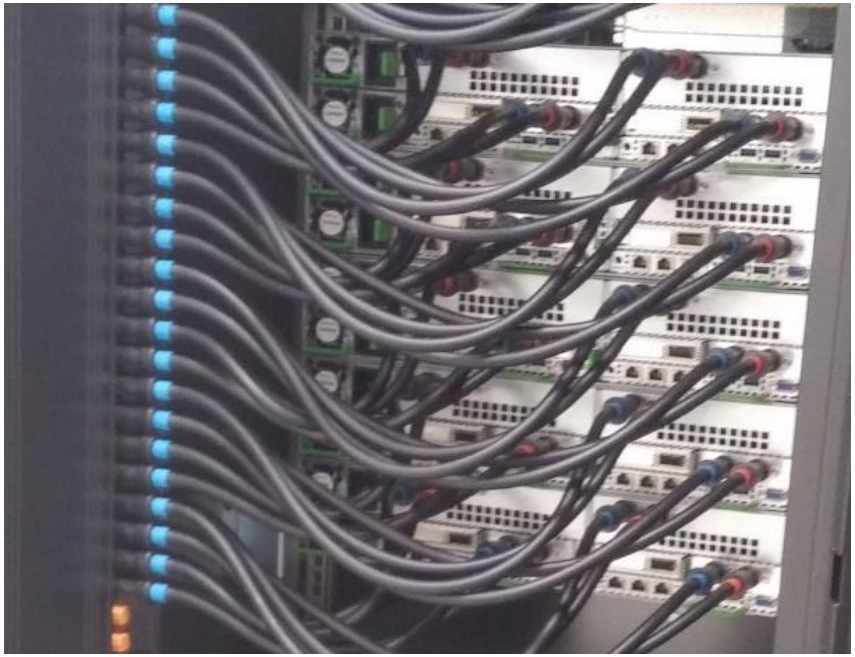
Liquid Cooling:
Warm water cooling direct-to-chip



L18S Warm water dry coolers & pumps

Direct-to-Chip Cooling Technology

Primergy CX400

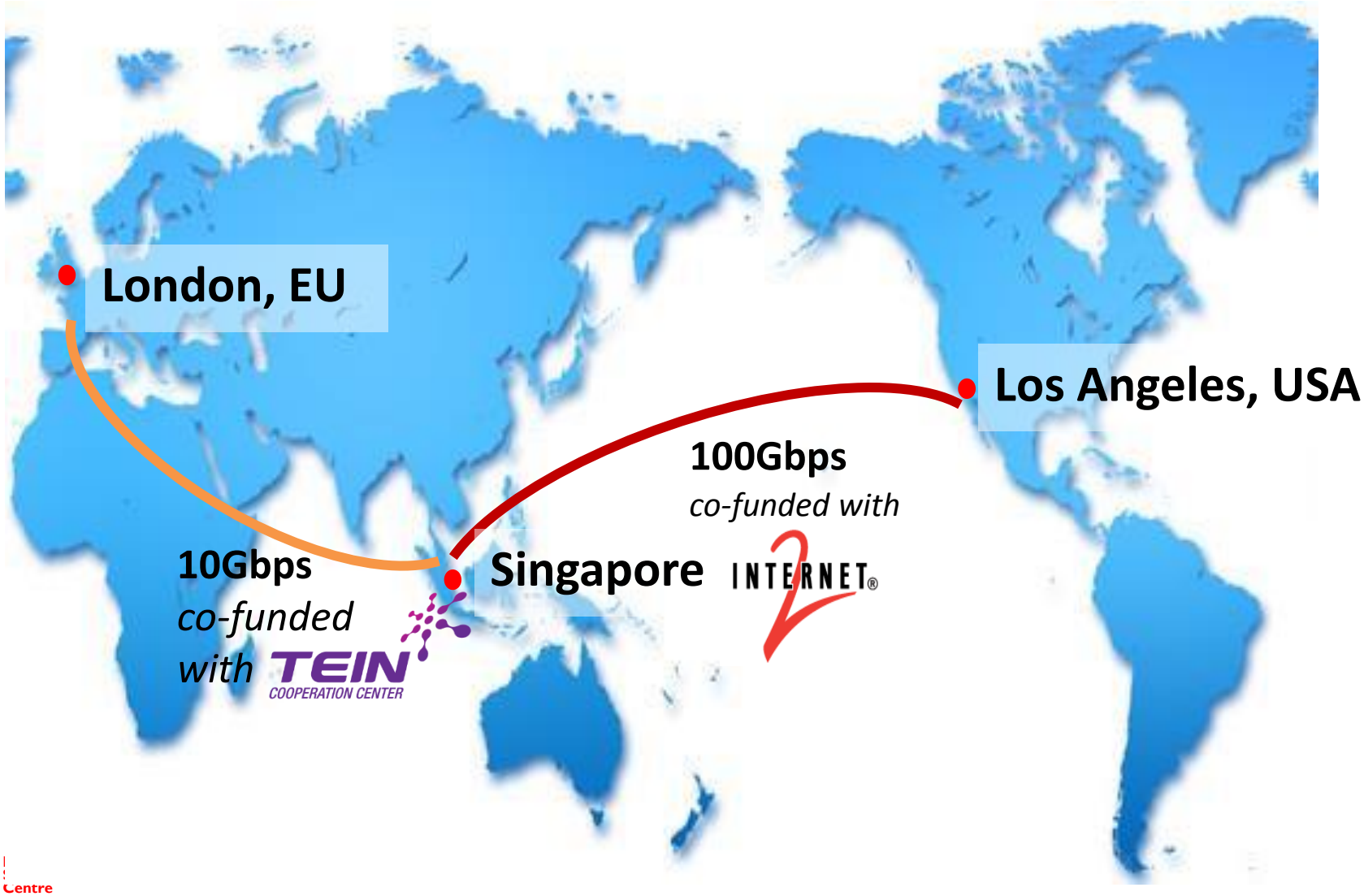


- **Direct-to-chip hot water (40 °C / 105 °F) based Cool-Central® Liquid Cooling captures between 60-80% of the servers heat.**
- **Helps to reduce data centre cooling costs by over 50% and allows for 2.5-5x higher data center density.**

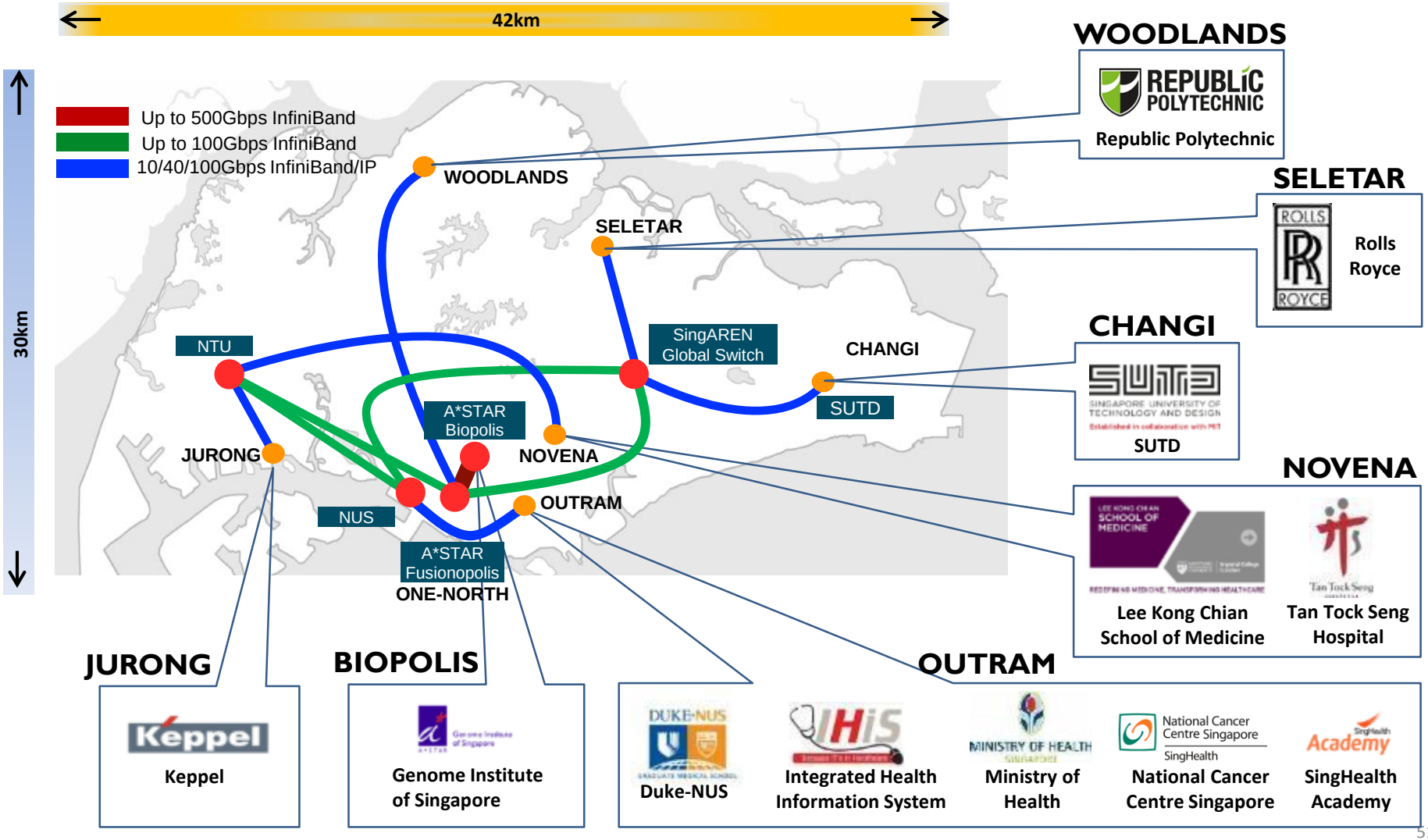
asetek

National Supercomputing Centre Singapore

NSCC co-funded International links



Envisaged High-Speed InfiniBand Fabric



National Supercomputing Centre Singapore

Application Areas

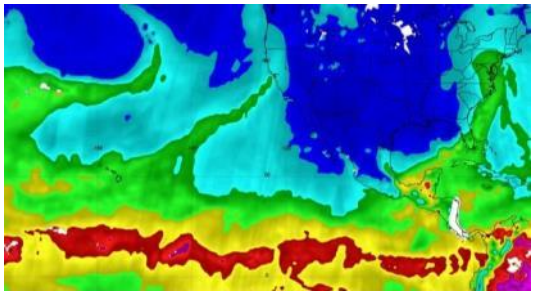
LIFE SCIENCES

Accelerate biomedical discoveries through **high performance applications in genomics**, thus improving the effectiveness of clinical treatments and personalised medicine.

[Image courtesy of insideHPC]

CLIMATE MODELLING

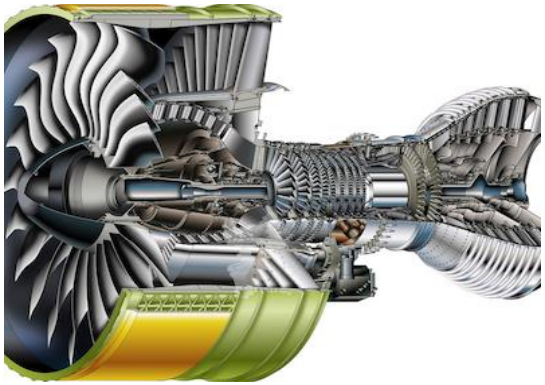
Contribute to atmospheric science and **improves the accuracy of weather forecasts** by broadening the range of parameters included in the simulations.



[Image courtesy of NASA]

MANUFACTURING

Enhance modeling, simulation and analysis to **speed up the design cycle for a faster time-to-market** for new and advanced products.



[Image courtesy of EnterpriseTech & Airbus]

Application Areas

COMPUTATIONAL FINANCE

Perform high performance **computational modelling** of **market conditions, pricing model, risk models**, and contingencies to allow financial institutions to **accurately meet real-time goals**.

[Image courtesy of MIR Labs]

DIGITAL MEDIA PRODUCTION

Accelerate **rendering with high realism**, reduces time to market for producers and increases the quality of production for users.

[Image courtesy of Omens Studios]

DATA CENTRE & NETWORKING

Offer an unprecedented high performance network testbed coupled with **high performance data analytics for quasi-real-time intrusion detection** and cybersecurity optimisation

https://www.nscc.sg

contact@nscc.sg

National Supercomputing Centre Singapore

RECOMMENDED: Training videos

About “real-world” data analysis

Videos

https://www.youtube.com/watch?v=F9ijhI86kGQ