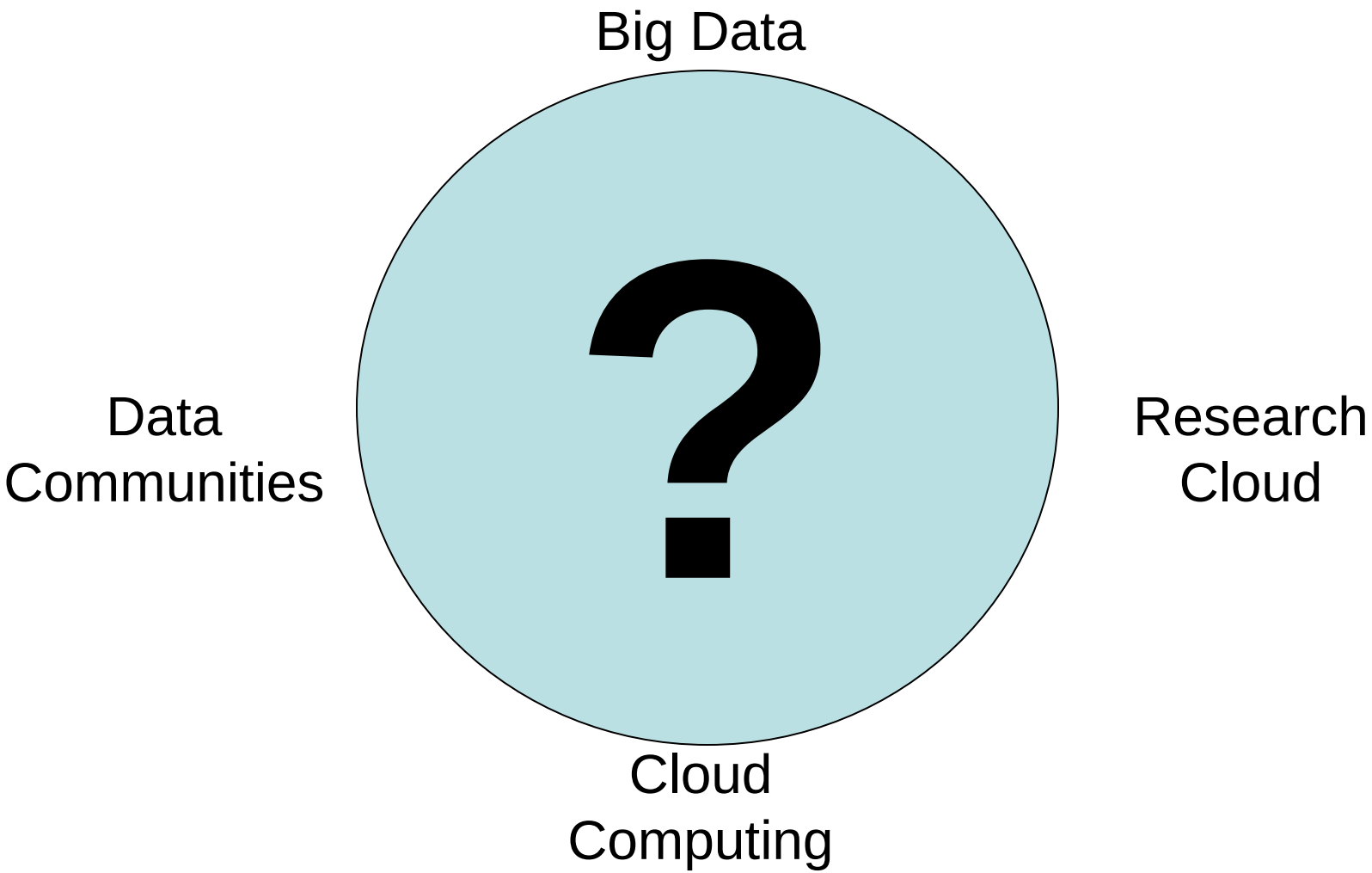
Research Cloud Data Communities

Bernard Meade,
Dr Steven Manos,
Prof Richard Sinnott,
A/Prof Christopher Fluke (Swinburne),
Dirk van der Knijff,
Dr Andy Tseng

What is it all about?



What is “Big Data”?

My definition:

“When accumulated data exceeds the capacity or capture rate of local resources, local storage and manipulation is impractical at best, impossible at worst.”

Examples of Big Data

- Human Genome
  - 100GB/personal human genome, 30,000 genomes processed in 2011
- Research Data Storage Infrastructure (RDSI)
  - Expected to exceed 100PB
- Large Hadron Collider (LHC)
  - 1TB/second, 13PB in 2010

Large Hadron Collider

In five minutes, the LHC will have filled 300 of these:

and 1,200 before the end of this talk.

Image source: http://www.westfordcomputer.com/service5.php

What is cloud computing?

"Computational resources offered as a platform, infrastructure or software service and made available on demand, with data storage and computation accessible via the Internet"

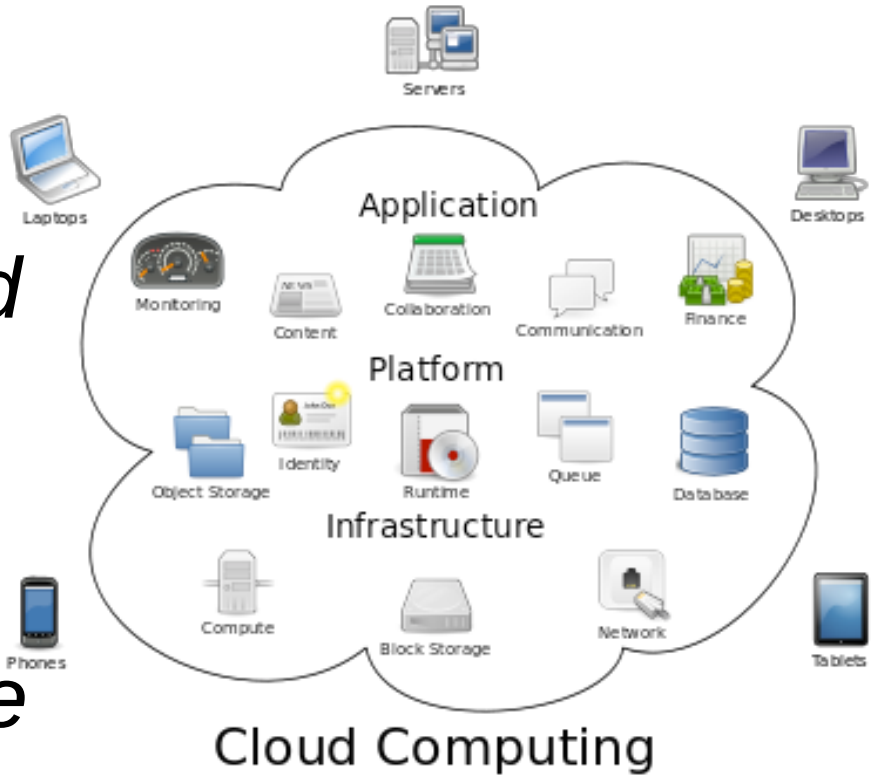


Image source: http://en.wikipedia.org/wiki/File:Cloud_computing.svg

What is cloud computing?

- Only pay for what you use
- Flexible load balancing
- Expansion on demand = cloud bursting
- “Cheap to Fail” means you can experiment
- US Federal Government’s “Cloud First” policy (2011)

Data Communities

- Research communities are the backbone of research
- Communities can form around disciplines, institutions, and even methodologies.
- Communities can also form around datasets and data collection resources and methods.

Data Communities

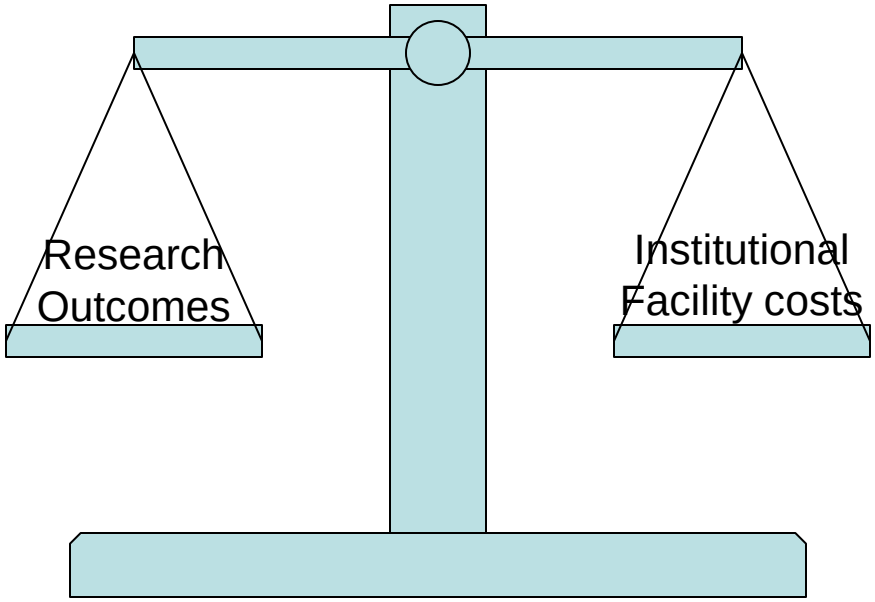
“Because the value of the data goes beyond the initial collection motivation, further research based on a dataset or collection of sets is brought about by community awareness.”

Data Communities

“This potential for reuse of data for entirely new research is a key ingredient to justifying expenditure on high-end resources, rather than myriad low-end resources.”

NeCTAR Research Cloud

- Launched in 2011
- 1,600+ research users
- 110 projects



Research Cloud Community Benefits

- Borders to collaboration are eased
- Free and Easy
- Sharing infrastructure
- Big Data infrastructure

Research Cloud Risks

- Ethical considerations
- Security management
- Network dependence
- Sustainability and technical capabilities

Cloud computing platforms for Astronomy

- “Data tsunami’ predicted over a decade ago
- Astronomy has been using HPC for many years
- HPC doesn’t suit all applications or researchers
- Learning curve for HPC is significant

Astronomy examples of Big Data

- Large Synoptic Survey Telescope
  - 20TB/night, 60PB over ten years
- Australian Square Kilometer Array Pathfinder (ASKAP)
  - 72TB/second (raw data stream)
  - 120 million Blu-ray discs/day
- Square Kilometer Array
  - ~1EB/day (2x daily global Internet traffic)
  - 100x LHC data collection

Astronomy examples of Big Data

Hubble Ultra Deep Field

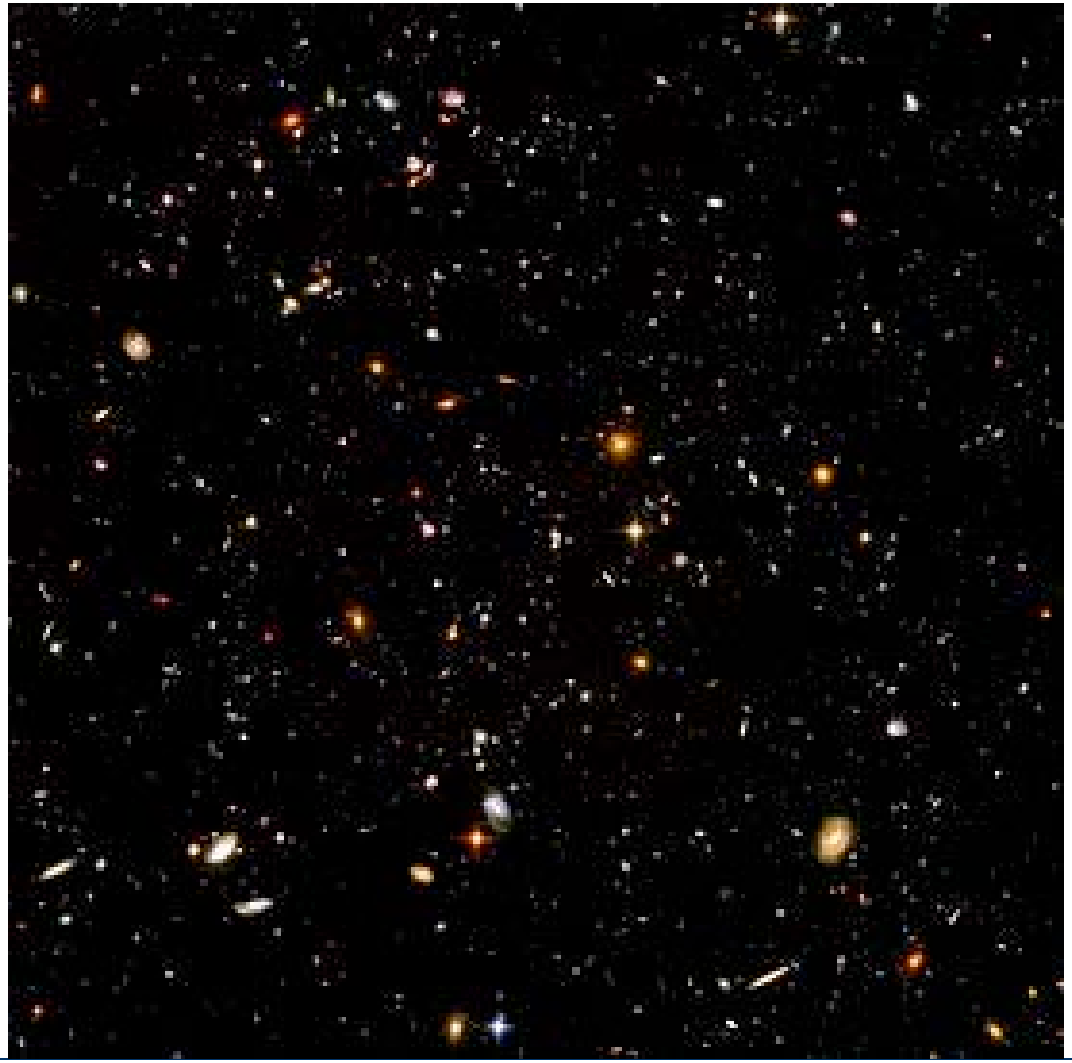


Image source: http://en.wikipedia.org/wiki/Hubble_Ultra-Deep_Field

Astronomy examples of Big Data

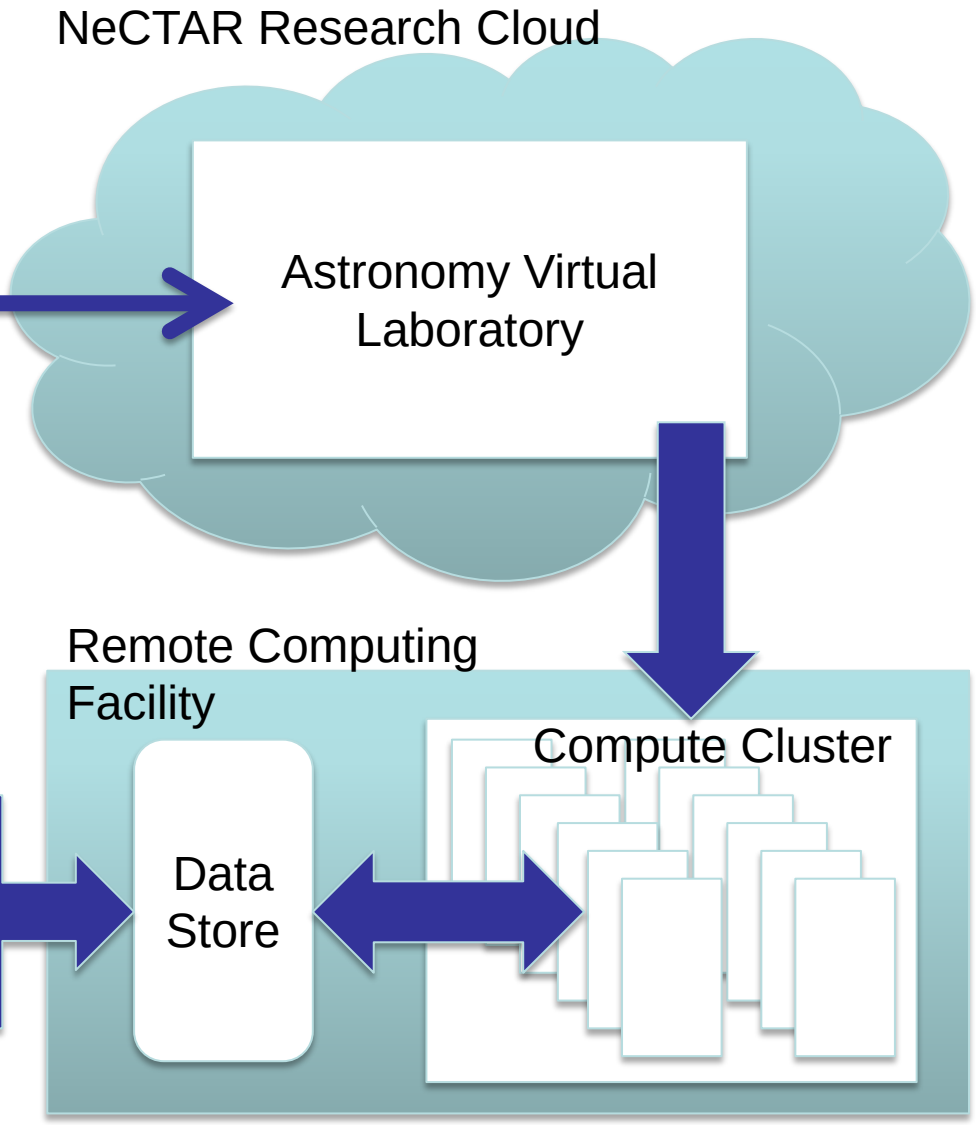
Hubble Ultra Deep Field

Hubble eXtreme Deep Field

Image source: http://en.wikipedia.org/wiki/Hubble_Extreme_Deep_Field

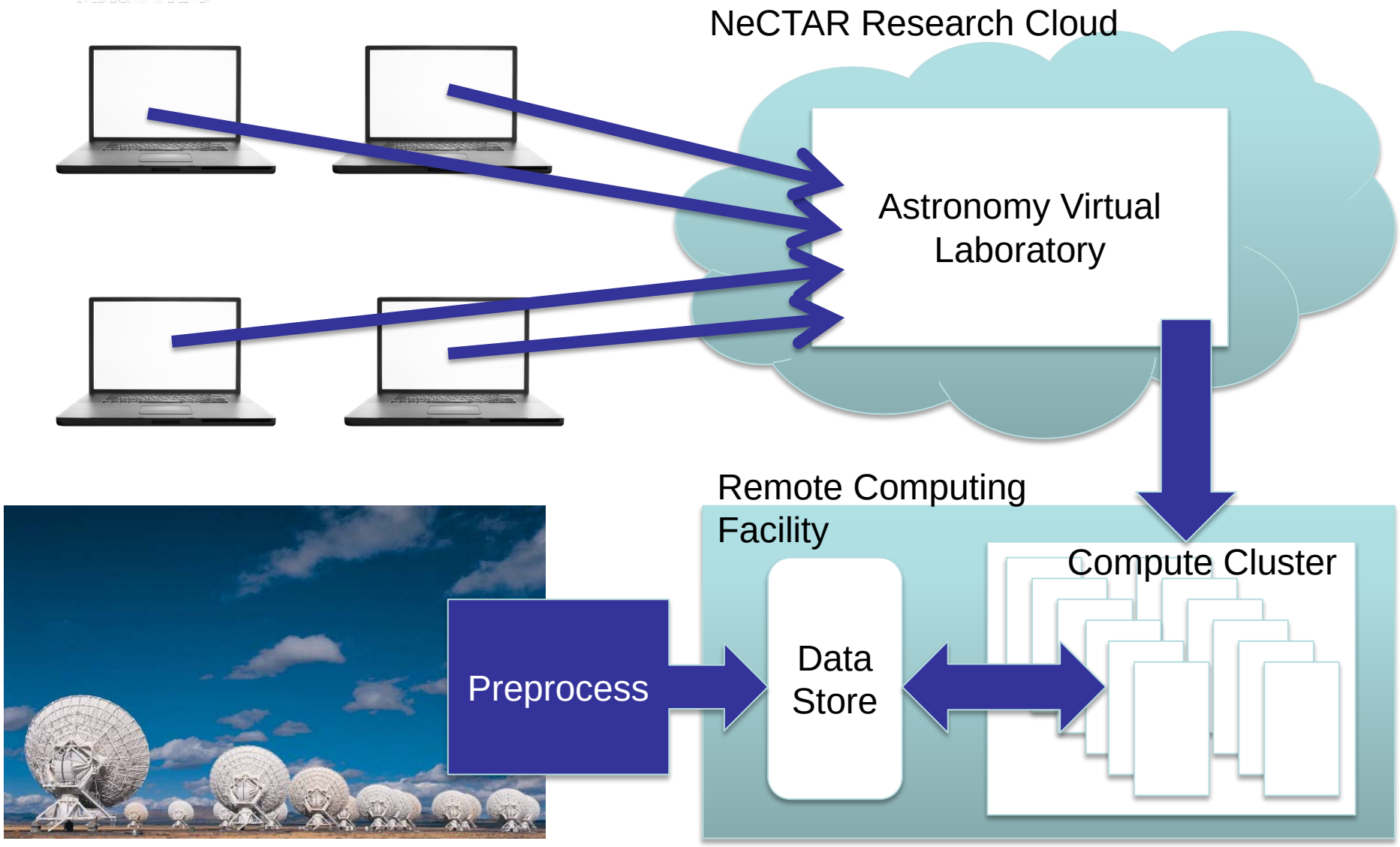
THE UNIVERSITY OF MELBOURNE

Possible cloud solution for Astronomy



THE UNIVERSITY OF MELBOURNE

Possible cloud solution for Astronomy



Cloud computing platforms

- More research papers are produced using archived data than from new data
- Reuse of data increases potential of research instruments
- VMs can be stored along with data for provenance and sharing
- Research building on a particular outcome can reuse a prebuilt VM

HPC and Research Cloud

- HPC systems are predominantly task-based
- HPC users guaranteed exclusive access to allocated resources for a limited time
- HPC jobs are queued
- HPC suited to well defined and bounded computation problems

- Cloud resources shared by being oversubscribed, but VMs are persistent
- Cloud is suited to ongoing continuous loads
- Cloud also has the capability to dynamically add VMs to cope with varying demand

Data management and provenance

- Research institutions need to develop policies for dealing with Big Data
- Reliability is fundamental
- Research data stored in the cloud can reduce the risk of data being lost or stolen
- Universities Modernization Fund assists UK Universities to take advantage of cloud

Data management and provenance

- Data provenance is critical to ensure integrity and quality of data
- Open Provenance Model provides guidelines on provenance information exchange

Cloud as part of an effective IT ecosystem

- Institutional IT is focused on consistency, reliability and cost saving
- Research necessarily challenges the way we do things
- Research IT must bridge the gap
- Research IT must cope with diverse disciplines, data formats, experience levels and expectations of quality and price

Cloud as part of an effective IT ecosystem

- Mapping a research problem to an IT problem is a major challenge
- Communities are key
- IT services need
  - Excellent communication
  - Community connections and training
  - Flexible underlying services that give control to the researcher

Where are we going?

- Cloud is here to stay
- Mobility devices connecting to cloud resources may replace the “desktop”
- E-Research => Research
- Utilities = electricity, water, lights, network
- Data collection will be designed to support many research activities

THE UNIVERSITY OF MELBOURNE

Thank you