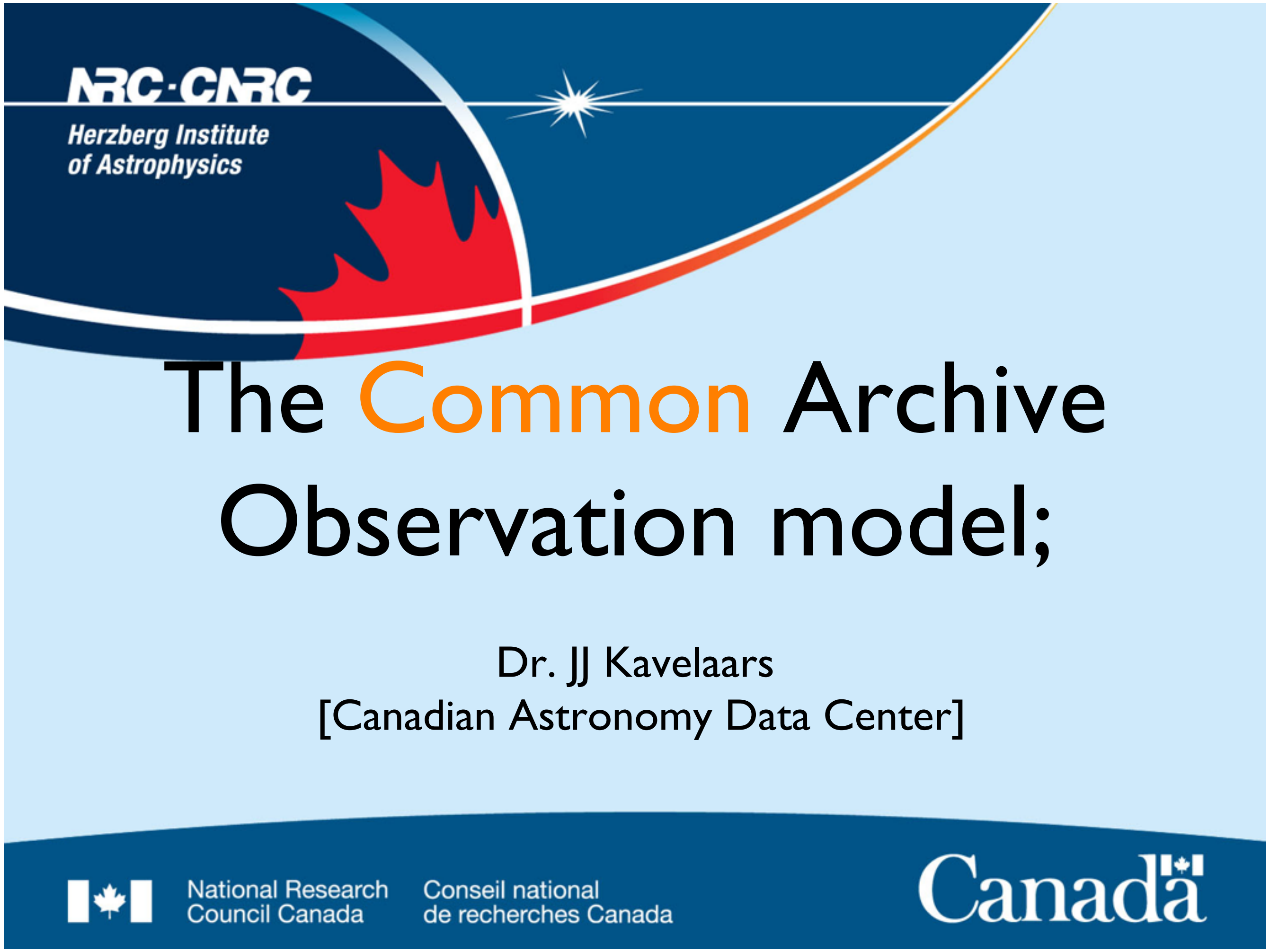
Keeping your data in the neighbourhood.

JJ Kavelaars

July 04, 2017

Canadian Astronomy Data Centre

Recalling APRIM 2008



NRC·CNRC

Herzberg Institute
of Astrophysics

The Common Archive Observation model;

Dr. JJ Kavelaars
[Canadian Astronomy Data Center]

National Research Council Canada

Conseil national de recherches Canada

Canada

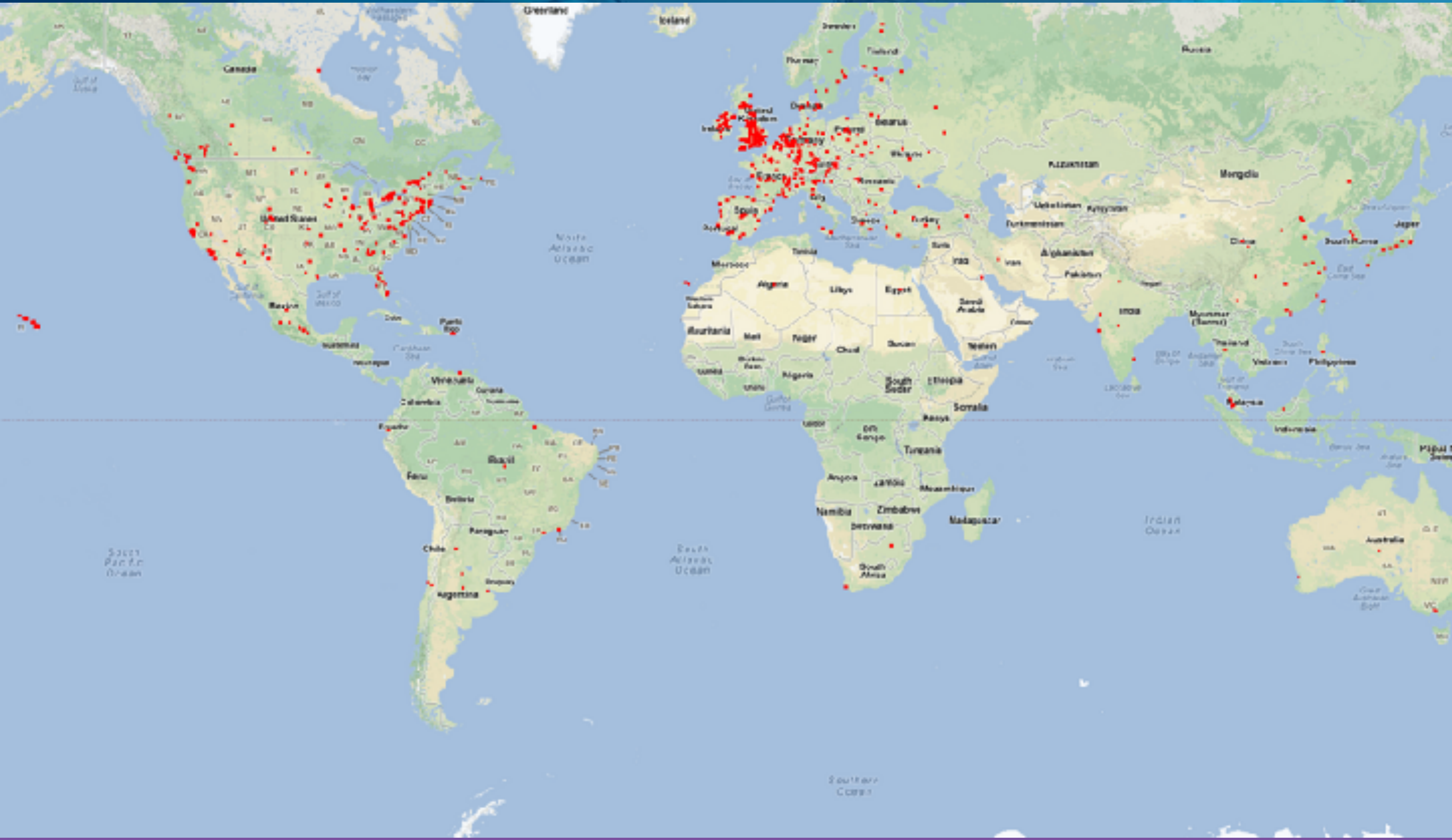
Canadian Astronomy Data Centre (CADC)

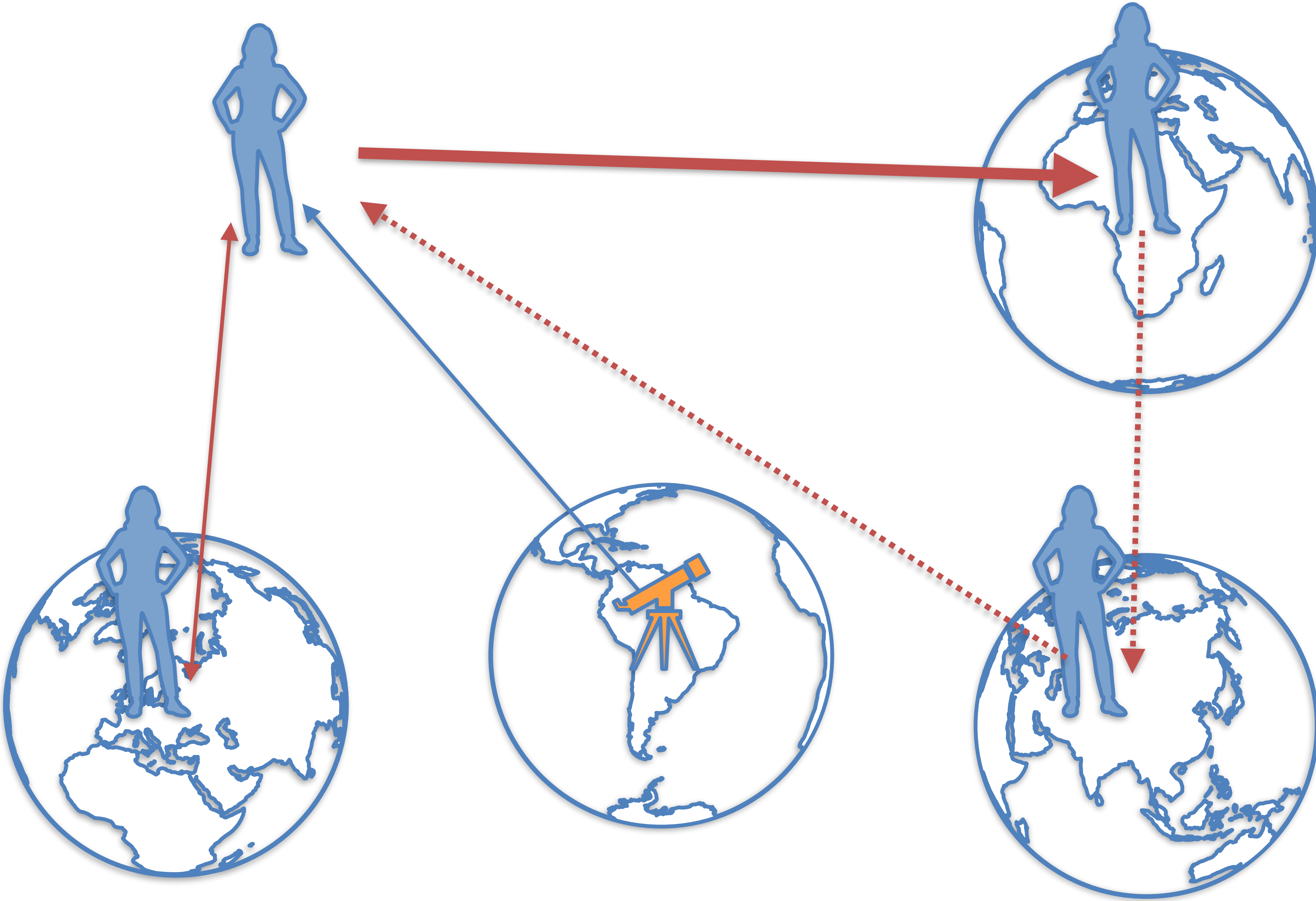
- Data archive housing collections from multiple (+10) telescopes:
  - Currently house about 2 PB of telescope archive data
- Development group working on standards in data archive system via the International Virtual Observatory Alliance (IVOA)
  - Common Archive Observation Model (CAOM and ObsCore)
  - VOSpace - Open storage system protocol.
  - Table Access Protocol (TAP) + Astronomy Data Query Language
  - Single-Sign-On Authentication and Access
- Lead development and support for the Canadian Advanced Network For Astronomical Research (CANFAR) - Cloud computing in Astronomy
- Research Astronomers investigating Dark Energy, Quasars, Galaxy Evolution, Stellar Atmosphere, the Trans Neptunian Region and Machine Learning in image and spectroscopy classification.

The current data problem for our users.

- Multiple PBs of data stored in archives and research repositories.
- Many research projects growing in the volume and complexity of data analysis required.
- Entire CADC archive delivered to users annually, over the internet
- Research partnerships are growing in size and becoming increasingly geographically distributed
- Visualization of complex datasets not obvious or straight forward.
- Resource requirements not always available near the expert.
- Data transfer overheads often exceed computation duration.

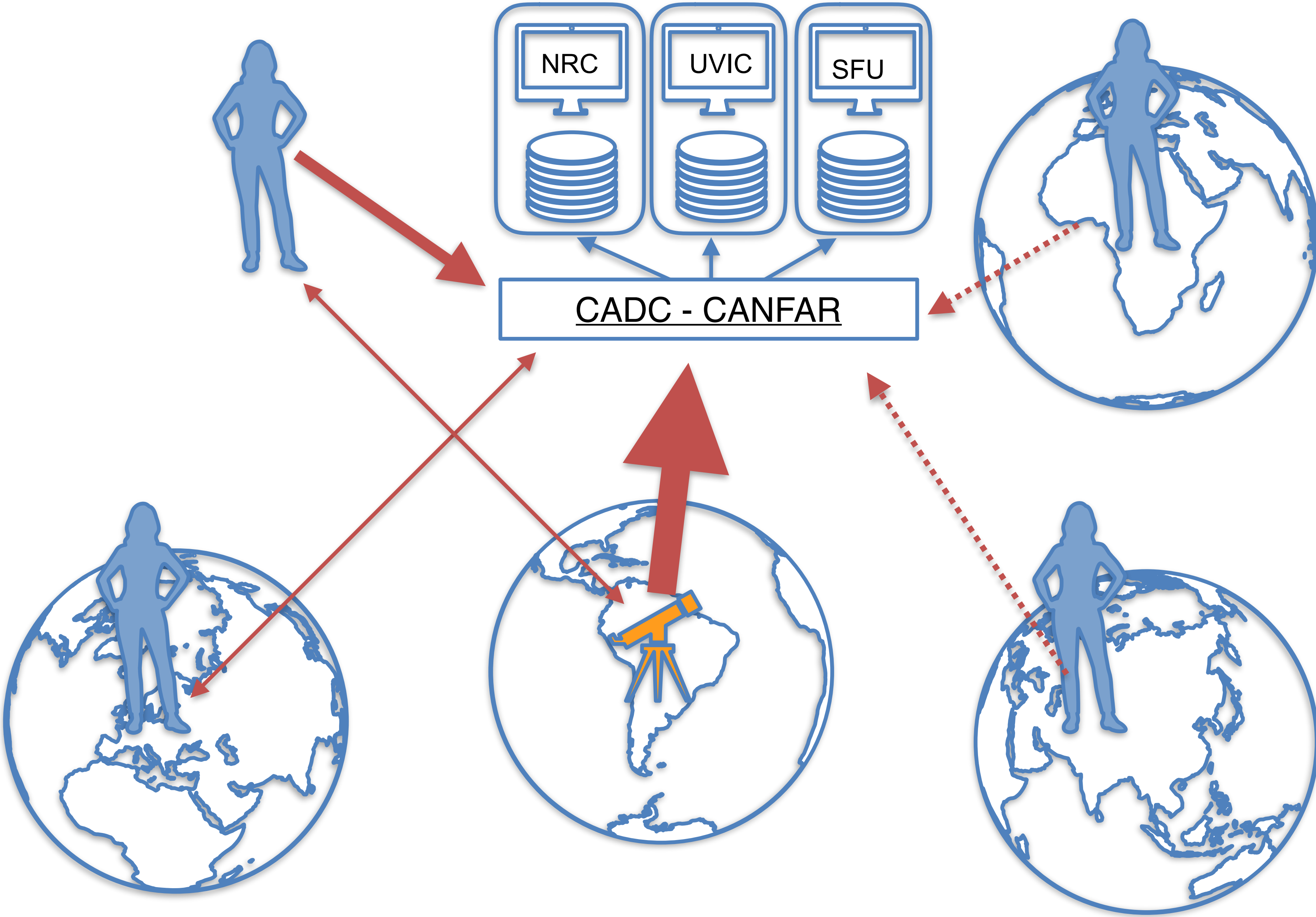
JCMT Data Delivery Location since 2006





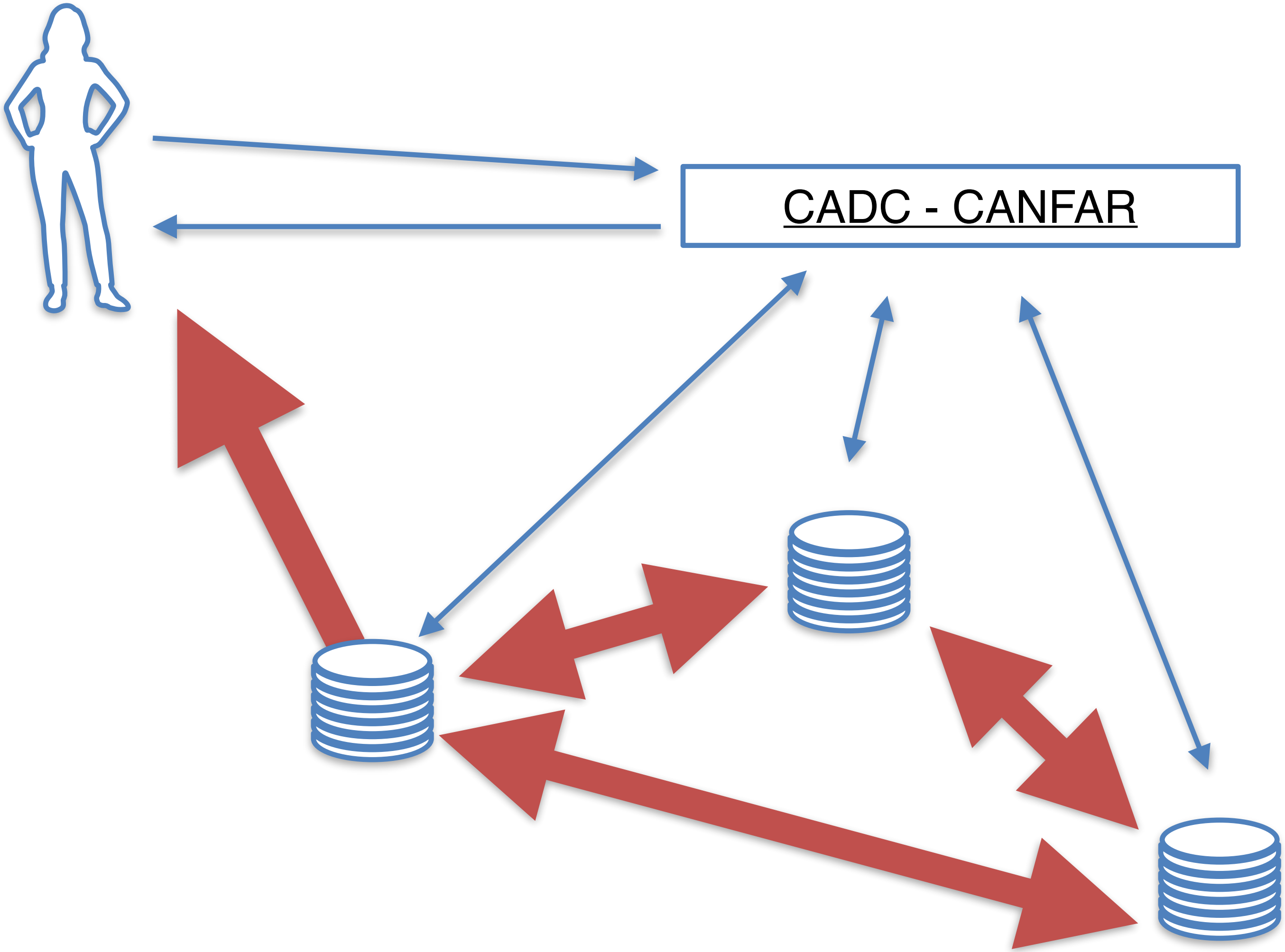
What is a data neighbourhood?

- The computing and data are in the same computing centre. Keep the data local to the compute.
- The ‘cloud’ abstracts your access to the data centre. When you run your computing on the cloud the jobs are moved to the computing centre that holds the data or the computing requests the data from the local storage. Should be transparent to users.
- CADC/CANFAR is currently distributed across three sites in British Columbia.
- About 1000 cores in a Cloud/VM system with about 4 PB of User and Archive storage.

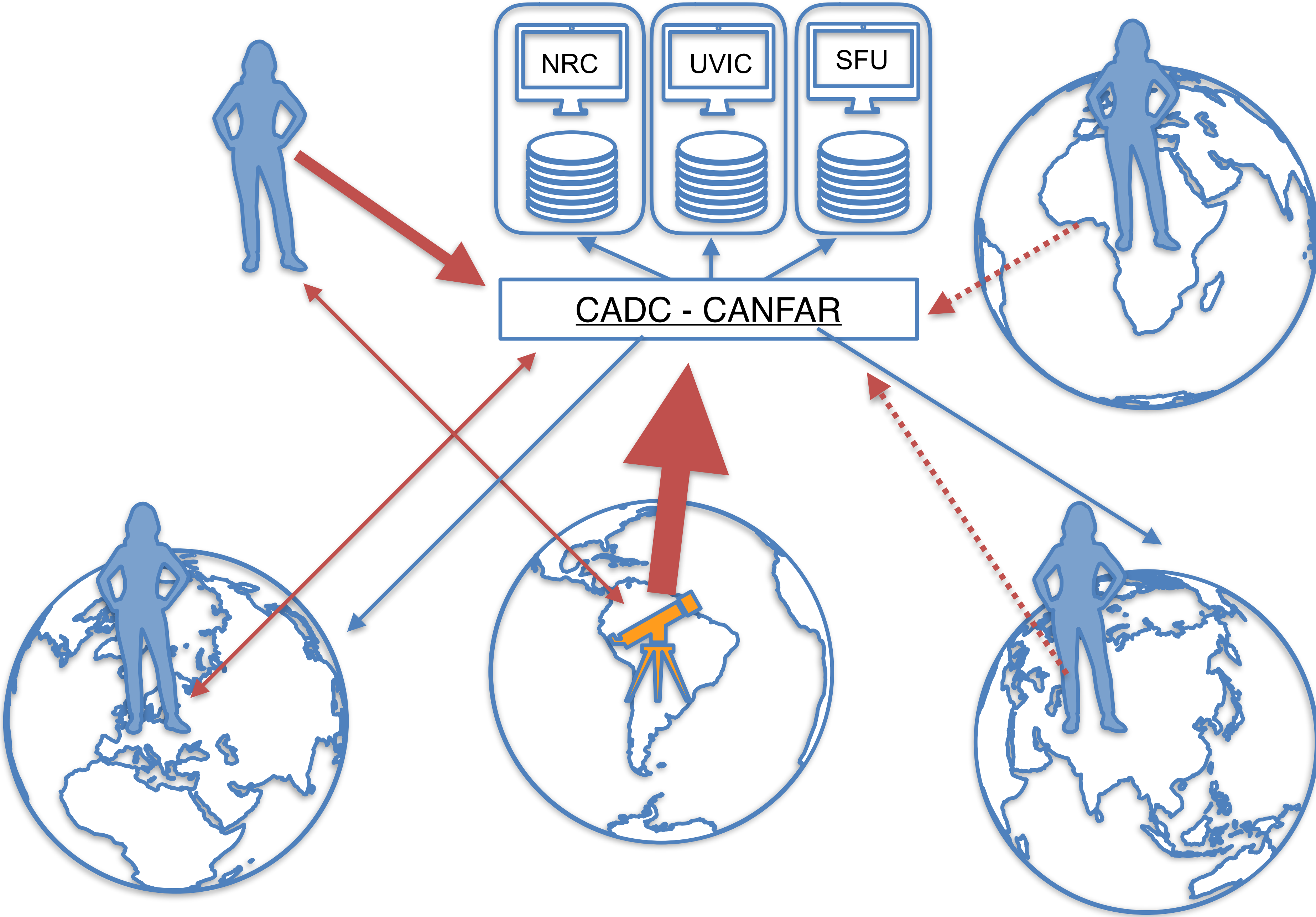

NRC
UVIC
SFU
CADC - CANFAR

Evolution of the distribution.

- On a global scale CANFAR / CADC is in a single geographic region
- Data are triplicated for storage security
- In house developed federation and storage
- Need for broader geographic distribution to improve user experience.
- Extend support for variable storage backends
- Exploring new federation approaches
- Eg. when you watch a YouTube video your request is redirected to a local version, often a copy that is store in a data centre located within the network of your ISP.
- Netflix abstracts this further, housing the data collection with the major ISPs and using DNS to control access via control of network.

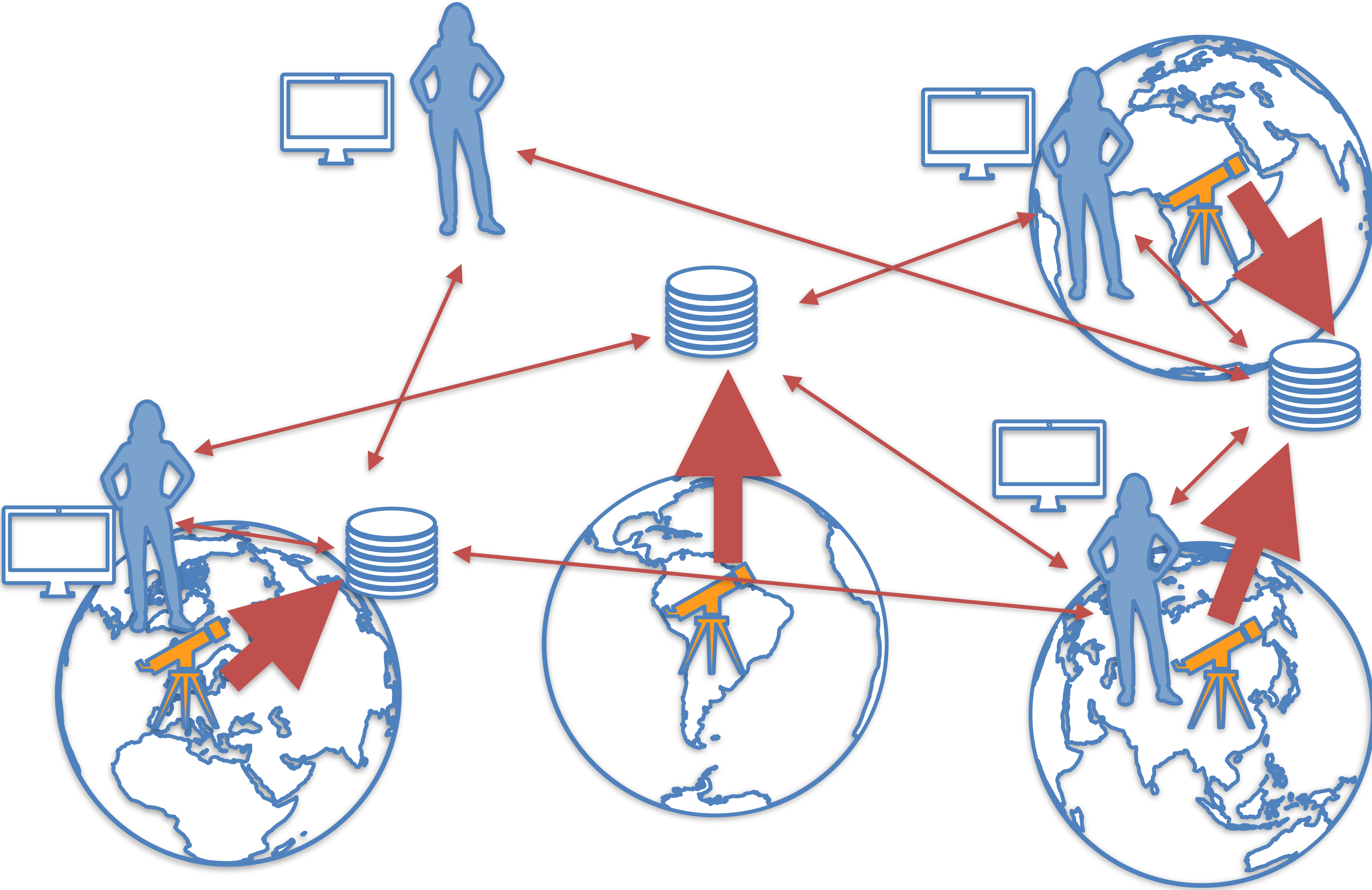


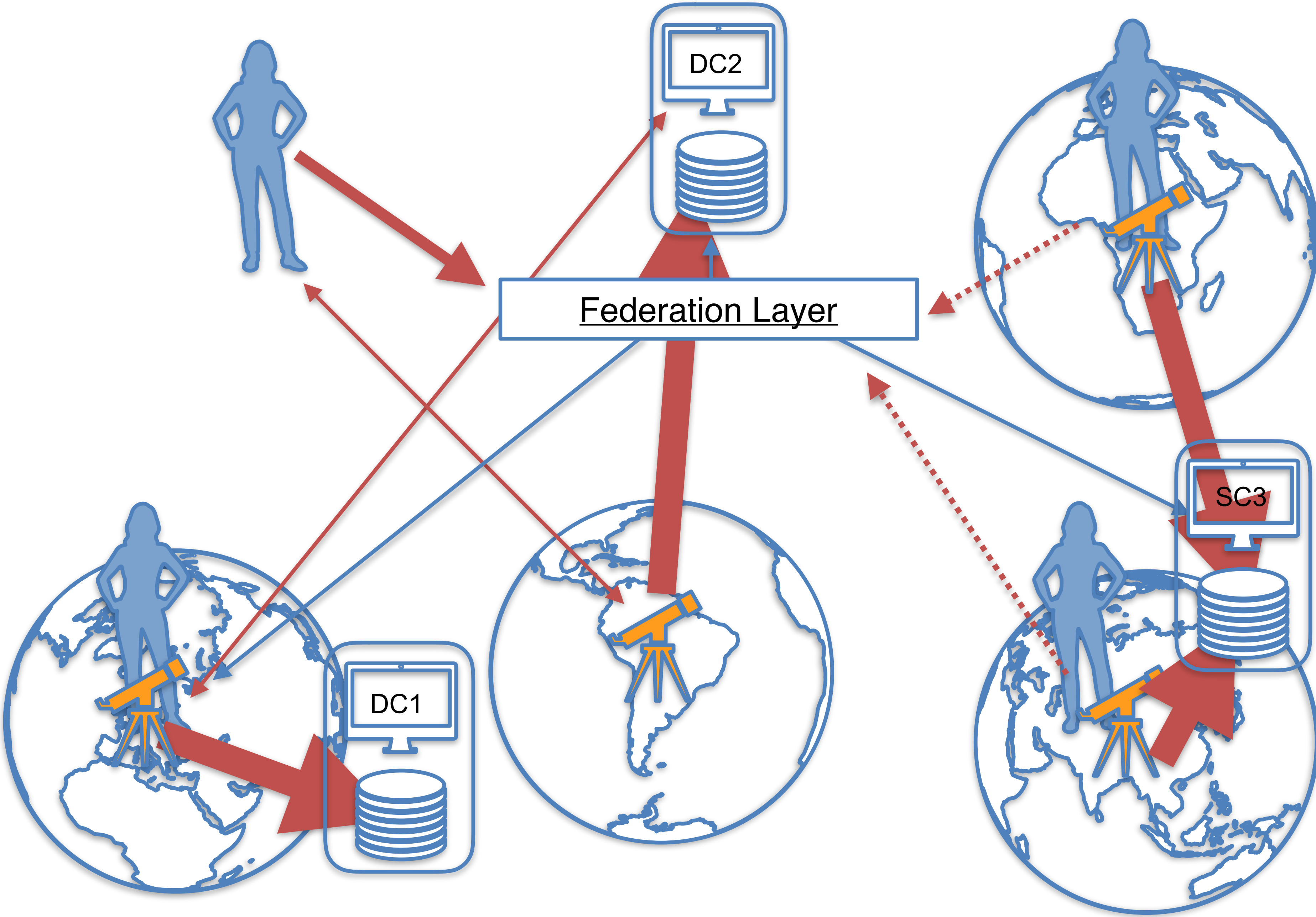
CADC - CANFAR


NRC
UVIC
SFU
CADC - CANFAR

A Neighbourhood storage system with a federated layer.

- Astronomers don’t control the Network, so can’t use Netflix style solution.
- Our goal is to share the resource not lock in a customer, requires standardize access layers - Open Standards approach via IVOA.
- Projects and telescopes are often multi-national organizations, users require global identities to access distributed datasets
- Resources provided by a variety of organizations which must cooperate to ensure access across organization boundaries, think open data like open skies
- Each partner contributes capacity for their complete dataset.




DC2
Federation Layer
SC3
DC1

Thank you

JJ Kavelaars
Canadian Astronomy Data Centre
Tel: 250-363-8694
JJ.Kavelaars@nrc-cnrc.gc.ca
www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca