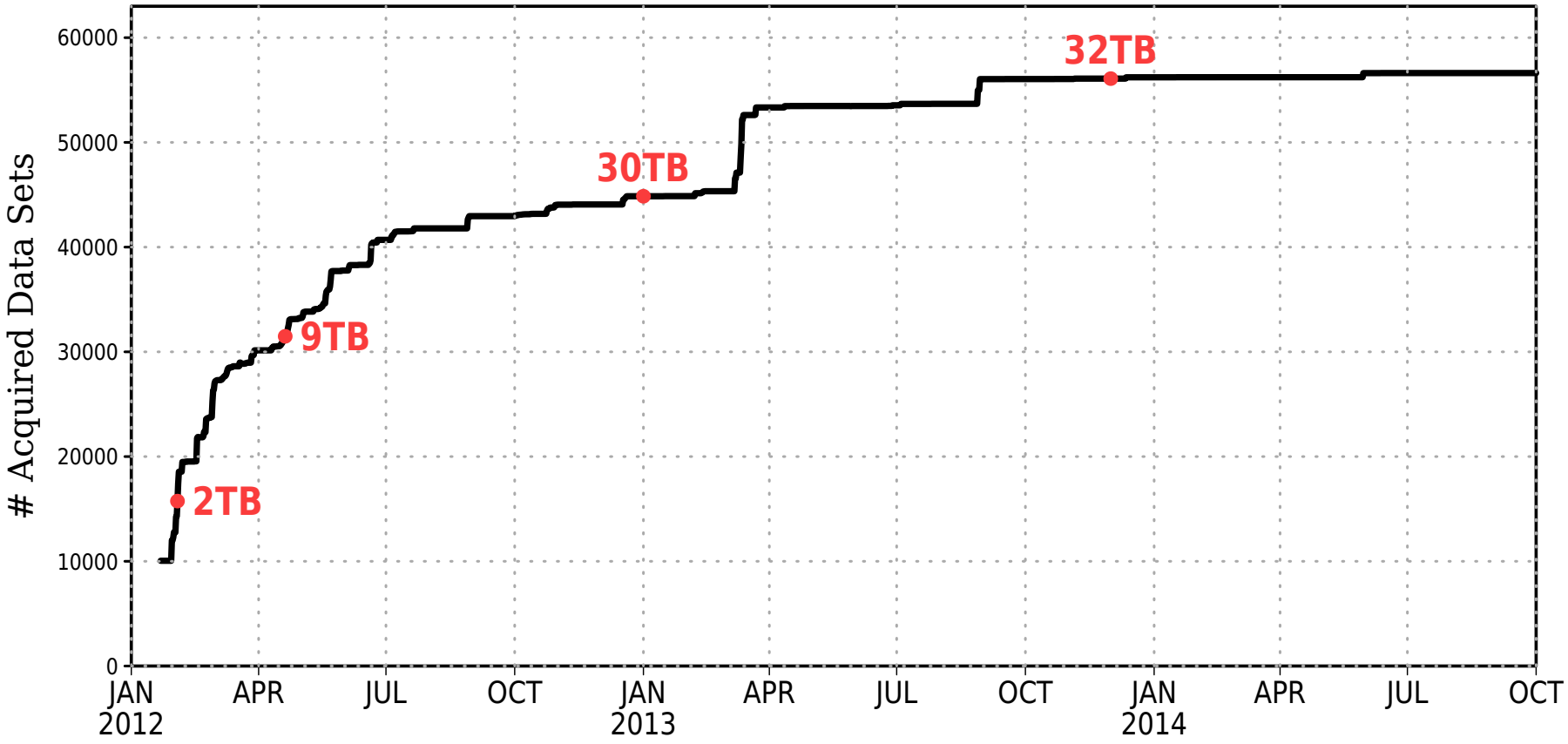
A User's Perspective on Acquisition and Management of CMIP5 Data

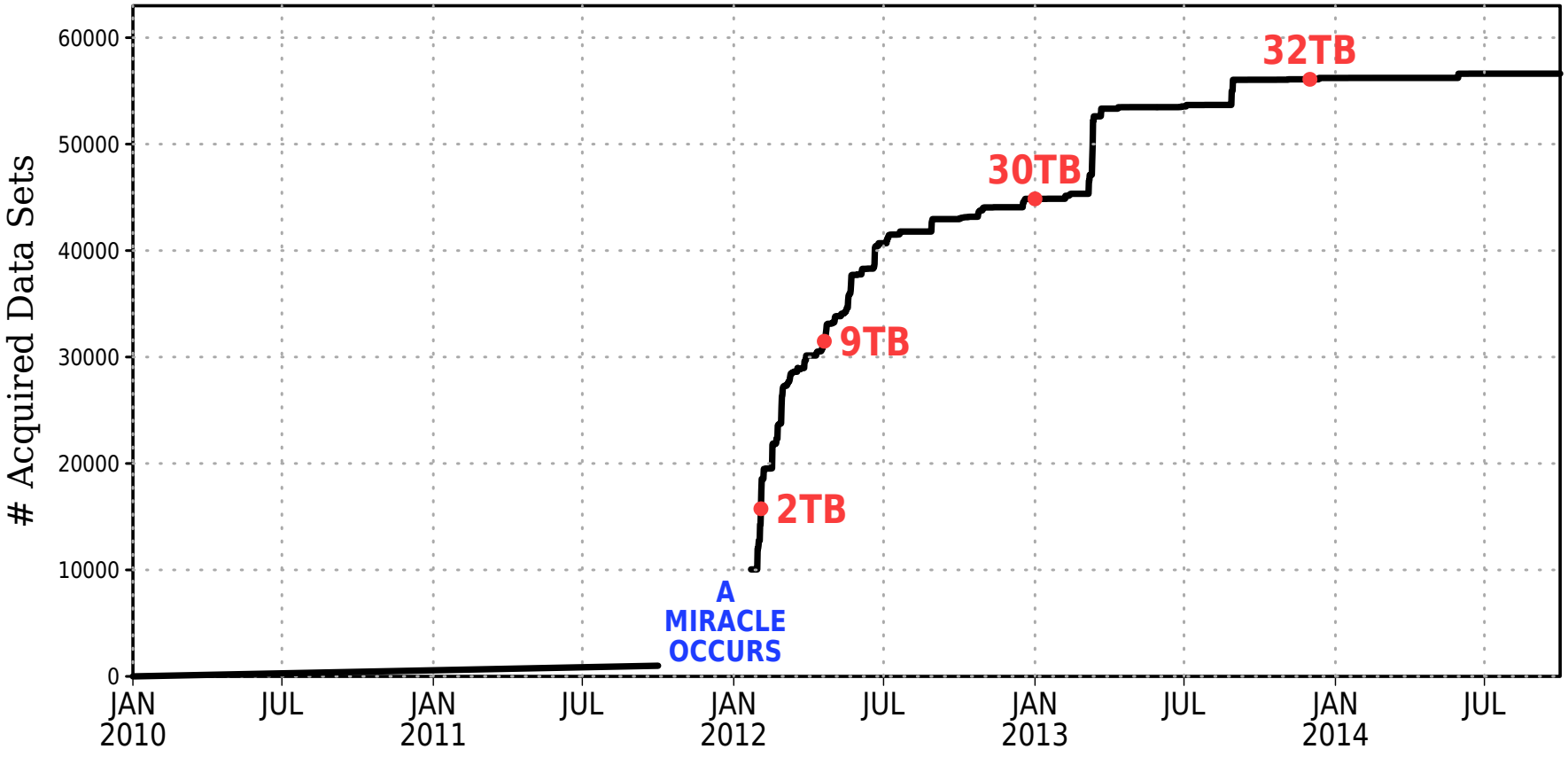
Jennifer Miletta Adams
George Mason University / COLA

ESGF2F, December 2014

COLA's CMIP5 Data Collection



COLA's CMIP5 Data Collection



Workflow Requirements

- No  ,  ,  ,  , et al.
- Script-Based
- Flexible
- Automated
- Runs in a UNIX environment

Workflow Elements

1. Create list of desired data:
   "All available models and ensembles for a subset of experiments, realms, frequencies, and variables"
2. Keep track of what has already been acquired
3. Identify what data are available
4. Get needed data
5. Make data user-friendly

Programmatic View of Workflow

```
while(1) {
  list(acquired);
  for(desired) {
    search(available);
    for(available) {
      if(!acquired) needed;
    }
    download(needed);
  }
}
```

Keep Track of Acquired Data

11 keywords are required:

```
cmip5
  /data
    /Experiment
      /Realm
        /Frequency
          /MIP-Table
            /Variable
              /Institute.Model
                /Ensemble
                  /Version
                    /datafiles.nc
```

Discovery of Available Data

Build a Dataset search URL:

http://pcmdi9.llnl.gov/esg-search/search?type=Dataset
&latest=true
&replica=false
&facets=id
&limit=0
&project=CMIP5
&experiment=piControl
&realm=atmos
&time_frequency=mon
&cmor_table=Amon
&variable=clt&variable=hfls....&variable=vas

Download Needed Data

1. Build a *file search URL* to determine number of files for each data set
2. Build a *wget URL* to download wget scripts; then give them unique names
3. Keep authentication certificates up-to-date
4. Monitor execution of wget scripts in a staging area
5. Put files in place under local directory structure

Make Data User-Friendly

- Create GrADS descriptor files
  - ✓ Aggregate files over time dimension
  - ✓ Make use of ensemble dimension when appropriate
  - ✓ Identify missing or overlapping time periods
  - ✓ Assign non-standard dimensions (e.g. basin averages)
  - ✓ Handle 365-day calendars
- Interpolate data on non-rectilinear grids
  - ✓ For ocean and sea ice realms
  - ✓ ESMF's RegridWeightGen generates the interpolation weights
  - ✓ Rotate vector fields from grid-relative to Earth-relative coordinates before interpolation

Complications	Solutions
Version number not with data	Retained during wget script acquisition
1000 File limit per wget script	***Please minimize file granularity!***
User authentication	Automated with MyProxyClient
Errors from wget	Never mind why, just keep trying. Failure *is* an option.
Some data nodes are friendlier than others	Data node blacklist
Missing or overlapping data	***DO NOT hide missing data with a non-linear time axis!***
Rotation of grid-relative vectors	***Please publish gridspec files!***
Data on wacky grids	ESMF's RegridWeightGen

Special thanks to:
Luca Cinquini, Estani Gonzalez, Gavin Bell, Lawson Hanson, and the CMIP5 Helpdesk!