



# Analysis Ready Data in Analytics

---

## Optimized Data Stores for Analysis of Big Earth Data in the Cloud

Christopher Lynnes, NASA/GSFC  
Hook Hua, NASA/JPL



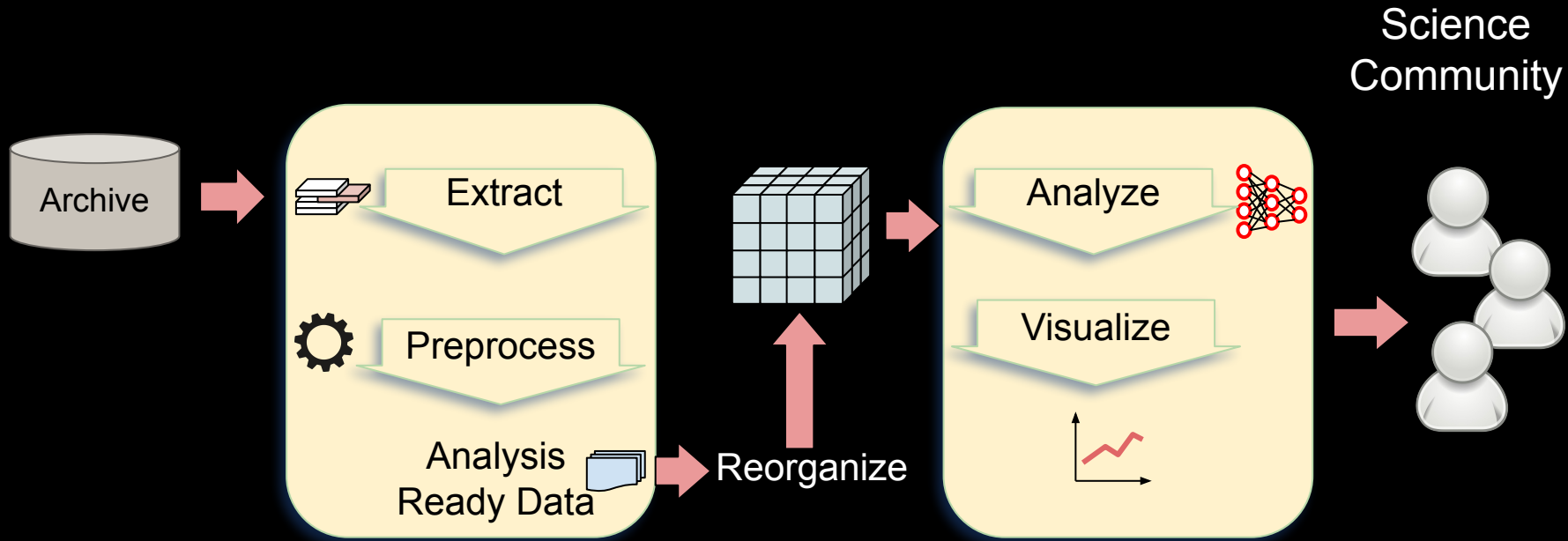
## **Analysis-Ready Data (ARD):**

Data that have been preprocessed to a form that allows immediate analysis by the end user

## **Analytics Optimized Data Store (AODS):**

A storage arrangement of data that supports fast analysis of Big Data

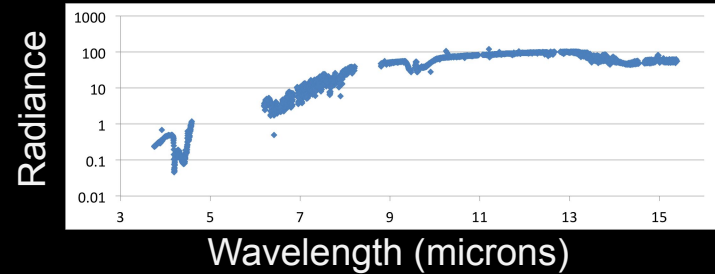
# Analysis-Ready Data vs. Analytics Optimized Data Stores



# ARD and NASA Data Processing Levels



**Level 1B - Calibrated + Geolocated**  
Calibrated radiance at a pixel

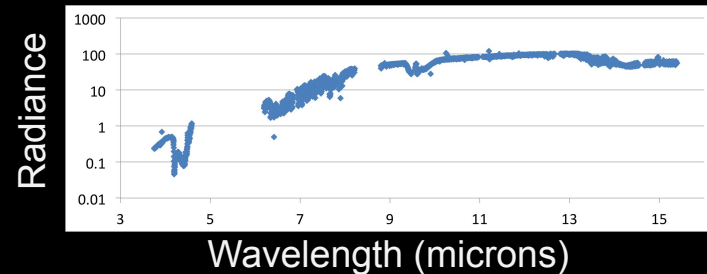


*Atmospheric Infrared Sounder (AIRS)  
data for 1 pixel on 2011-08-11*

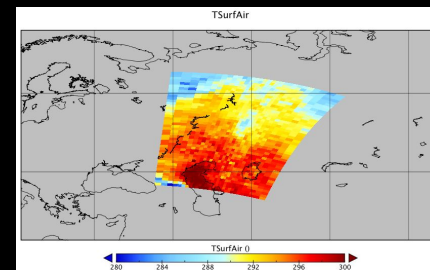
# ARD and NASA Data Processing Levels



**Level 1B - Calibrated + Geolocated**  
Calibrated radiance at a pixel



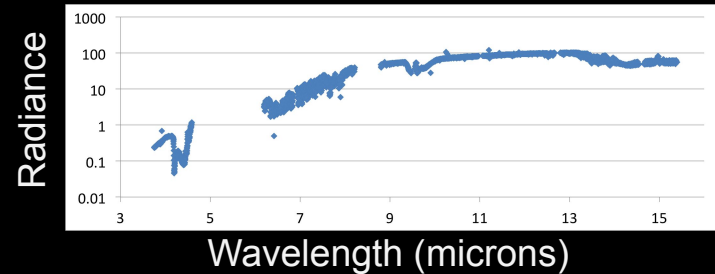
**Level 2 - Geophysical Parameter**  
E.g., Surface air temperature for one scene



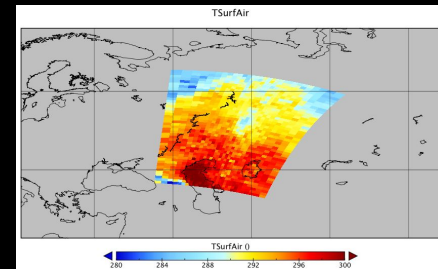
# ARD and NASA Data Processing Levels



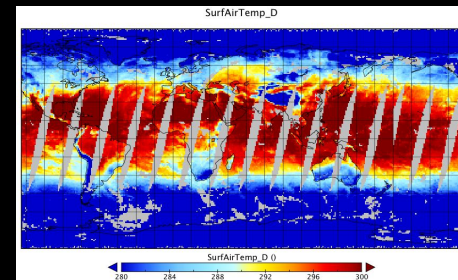
**Level 1B - Calibrated + Geolocated**  
Calibrated radiance at a pixel



**Level 2 - Geophysical Parameter**  
E.g., Surface air temperature for one scene

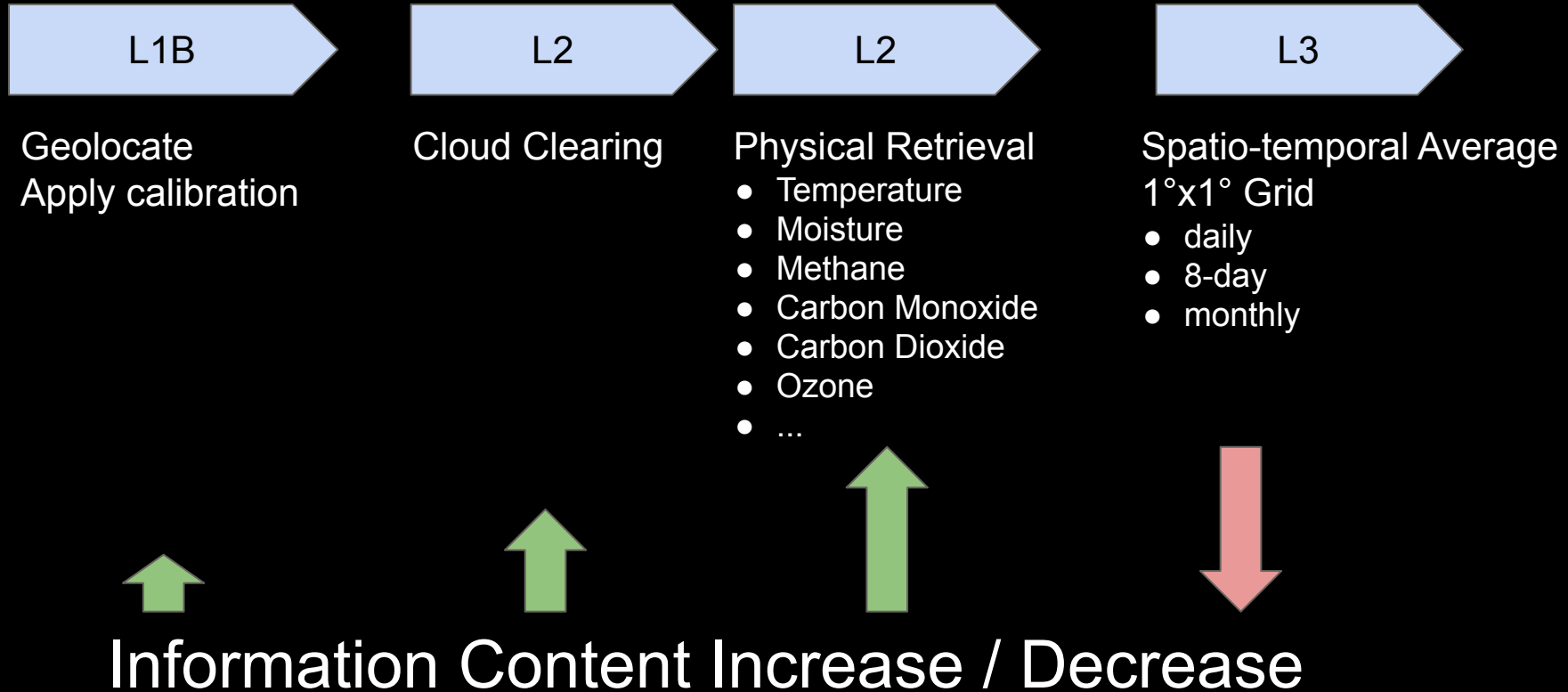


**Level 3 - Spatio-temporally Averaged**  
E.g., Global surface air temperature for one night  
(Almost-global coverage, rectilinear grid)





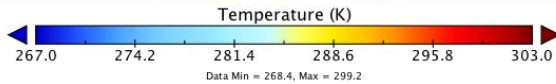
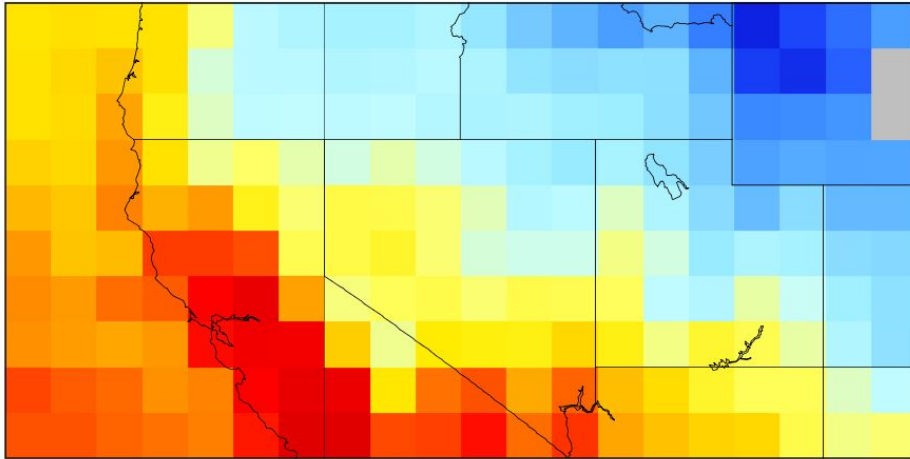
# Analysis-Ready AIRS Data



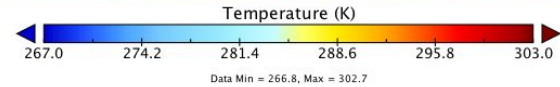
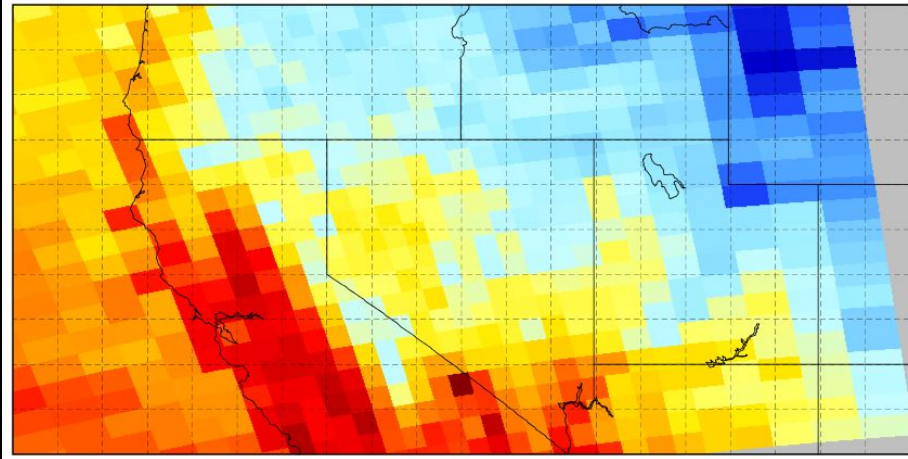
# Caveats regarding gridded (L3) data



Level 3 Surface Air Temperature



Level 2 Surface Air Temperature



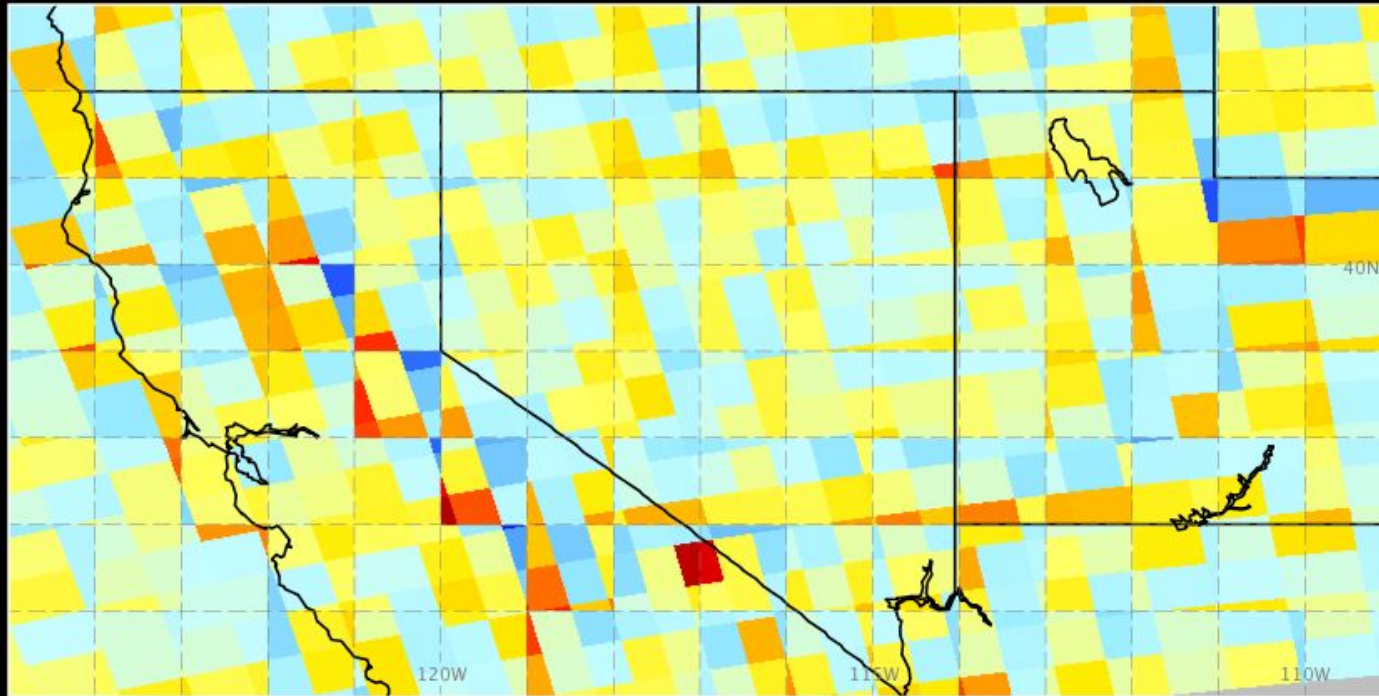
*Sacrificing spatial resolution for convenience*



# Caveats regarding gridded (L3) data



Surface Air Temperature Difference: Level 2 – Level 3

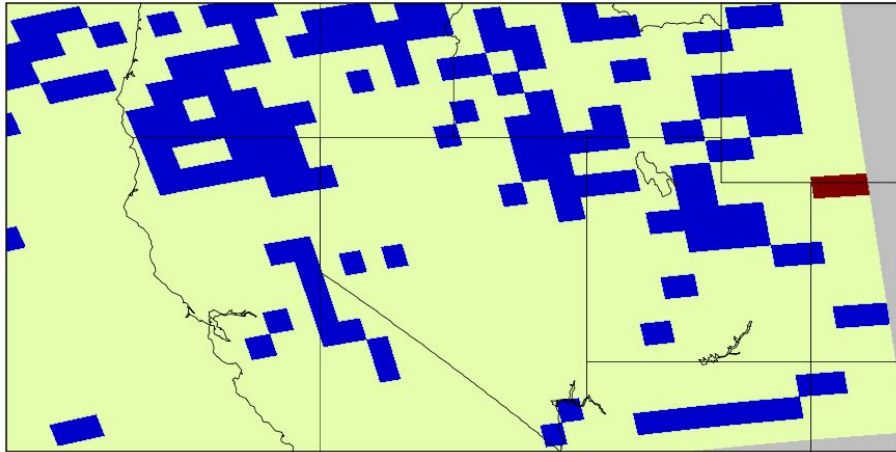


Data Min = -9, Max = 11

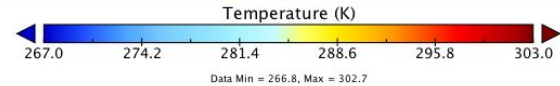
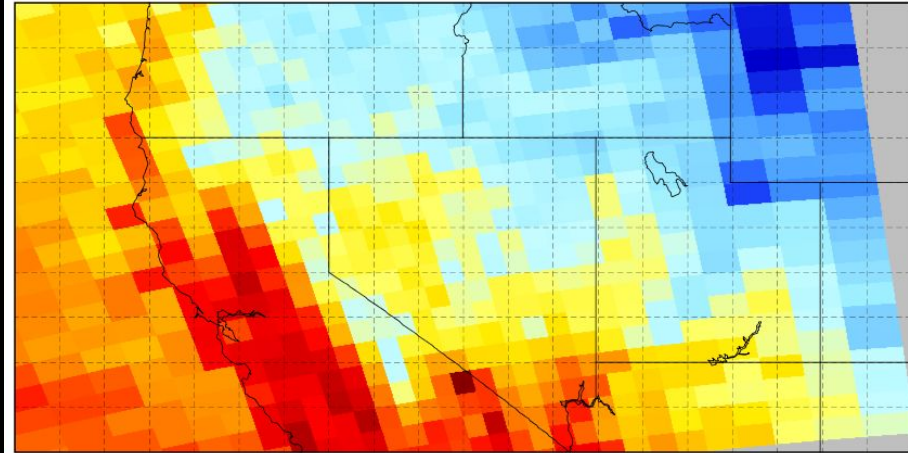
# Caveats regarding gridded (L3) data



Level 2 Surface Air Temperature – Quality Control Flag



Level 2 Surface Air Temperature

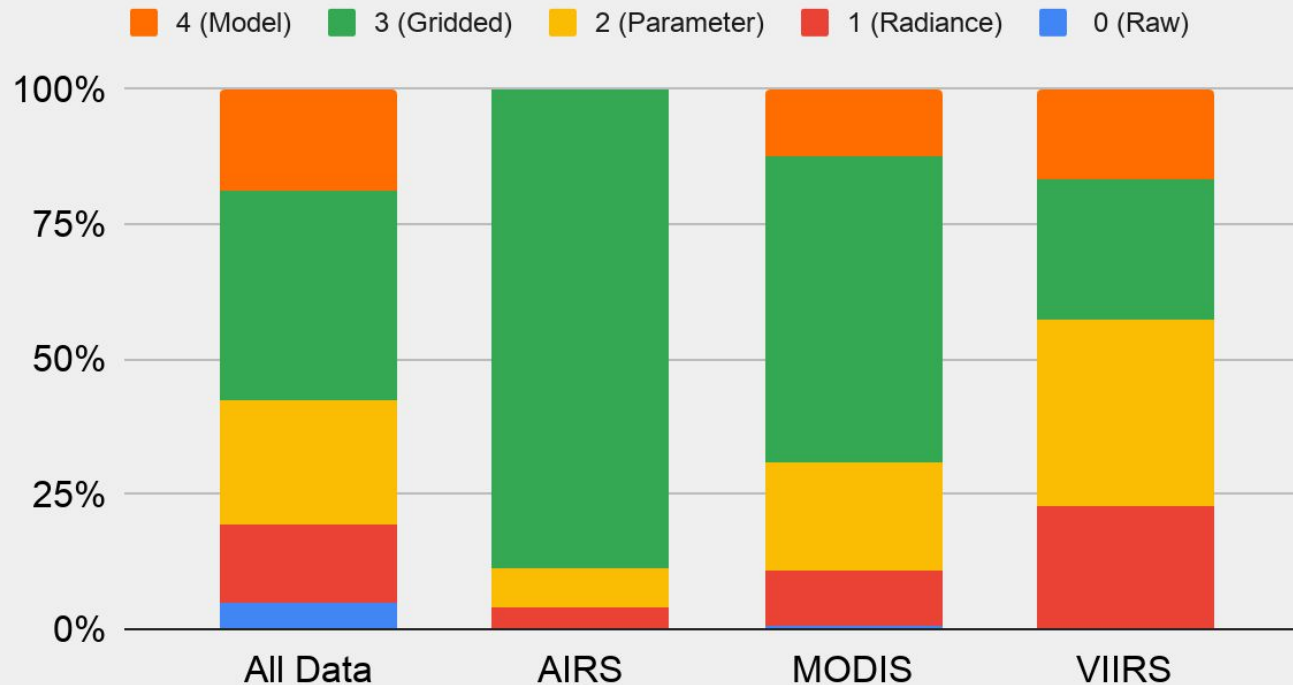


*Forgoing control over quality flag handling*

# Yet: Level 3 (Gridded) data is the most popular



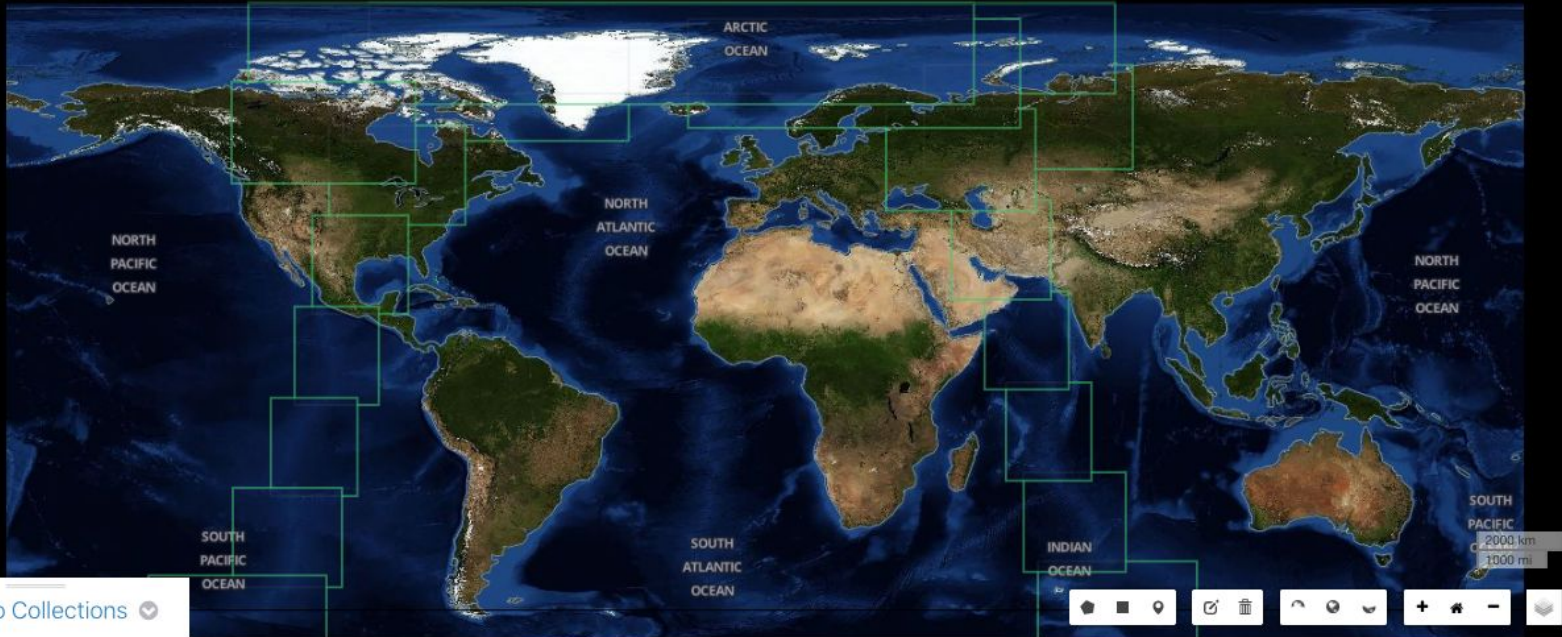
## Percentage of Users Acquiring Data by Processing Level





*Just how inconvenient ARE Level 2 Swaths???*

# Example: Surface Temperature Trends over the Globe



AIRS/Aqua L2 Standard Physical Retrieval (AIRS-only) V006 (AIRS2RET) at GES DISC [View Details](#)

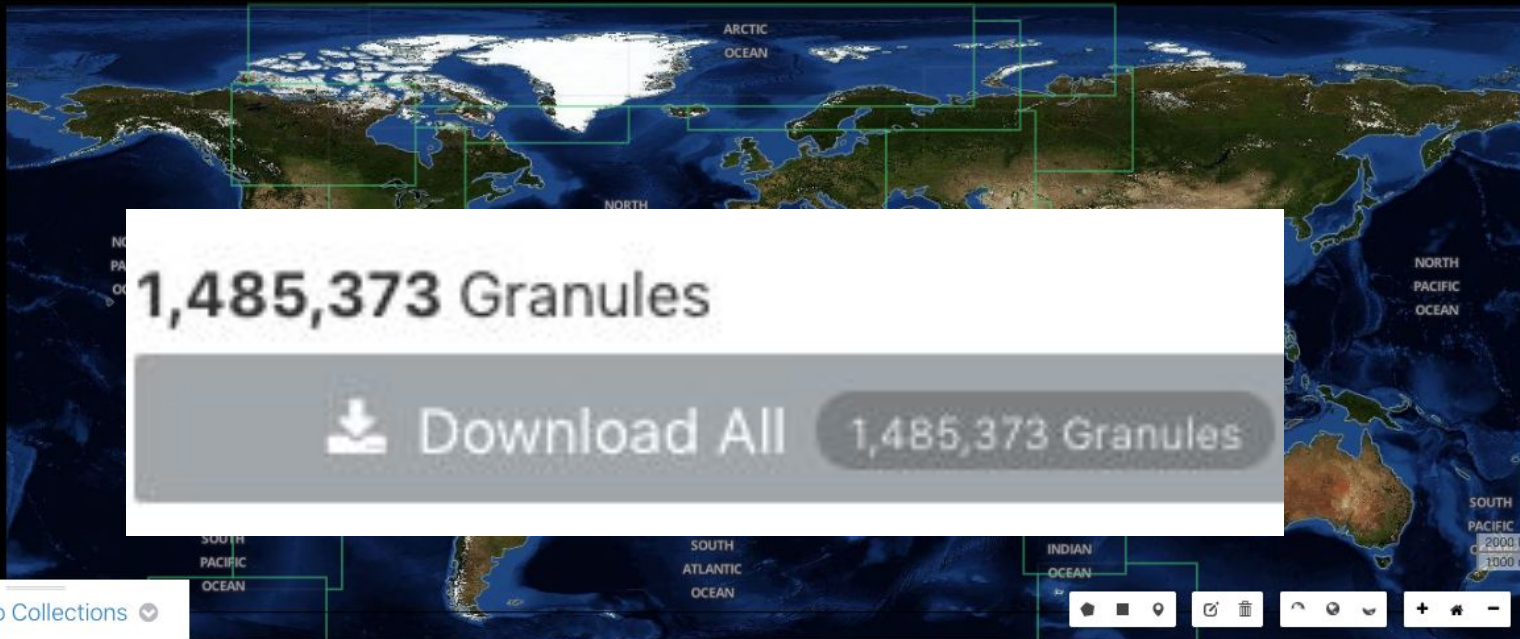
Sort by: Start Date, Newest First  Granule Search:  [Granule Filters](#)

1,485,373 Granules

[Download All](#) 1,485,373 Granules



# Example: Surface Temperature Trends over the Globe



[Back to Collections](#)

**AIRS/Aqua L2 Standard Physical Retrieval (AIRS-only) V006 (AIRS2RET) at GES DISC** [View Details](#)

Sort by: Start Date, Newest First  Granule Search:  [Granule Filters](#)

1,485,373 Granules

[Download All](#) 1,485,373 Granules

# Example: Surface Temperature Trends over the ~~Globe~~ Continental United States



Rectangle: SW: 24.97033, -124.89381 NE: 50.24074, -64.89502

103,772 Granules + Add to

Download All 103,772 Granules

[Back to Collections](#)

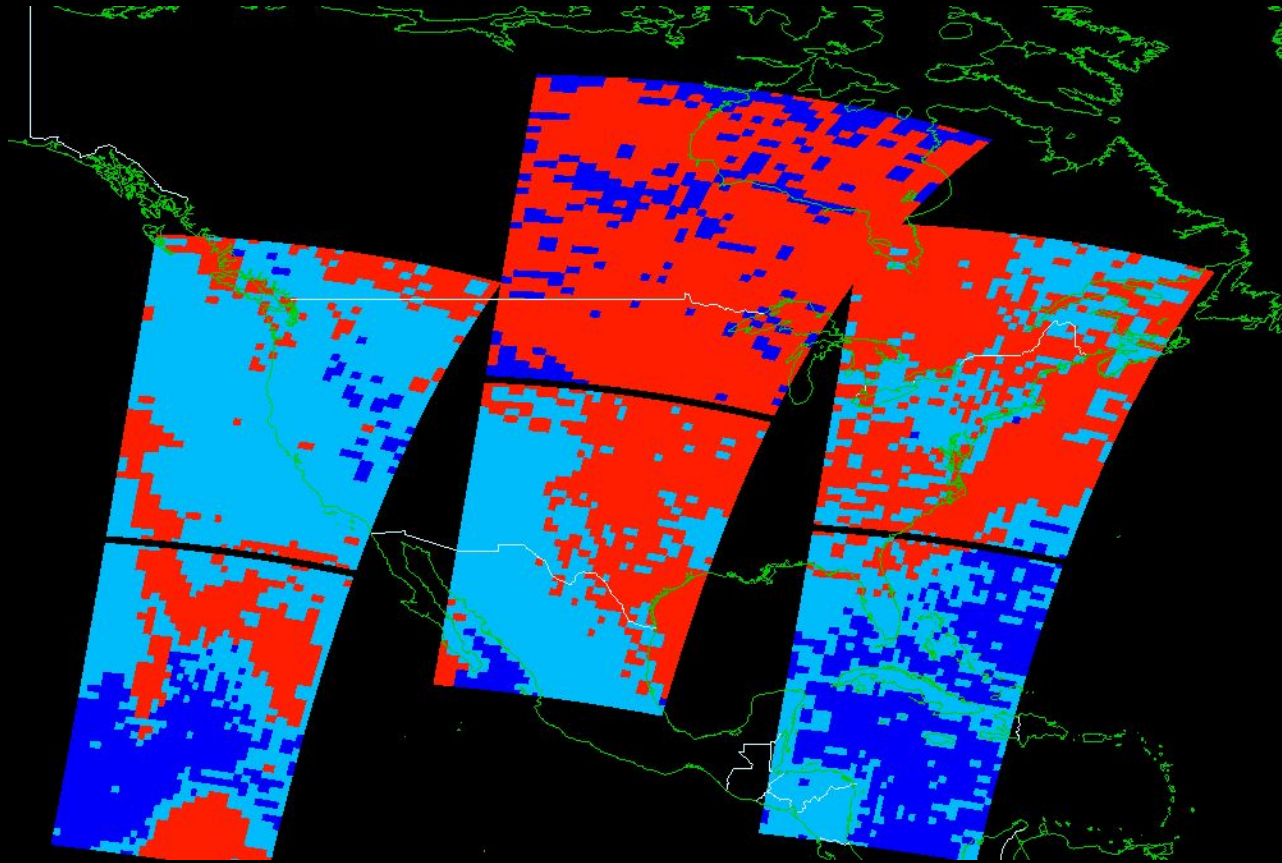
**AIRS/Aqua L2 Standard Physical Retrieval (AIRS-only) V006 (AIRS2RET) at GES DISC** [View Details](#)

Sort by: Start Date, Newest First Granule Search: Search Single of Multiple Granule IDs... [Granule Filters](#)

103,772 Granules + Add to project

Download All 103,772 Granules

Assembling proper pixels is no picnic either...

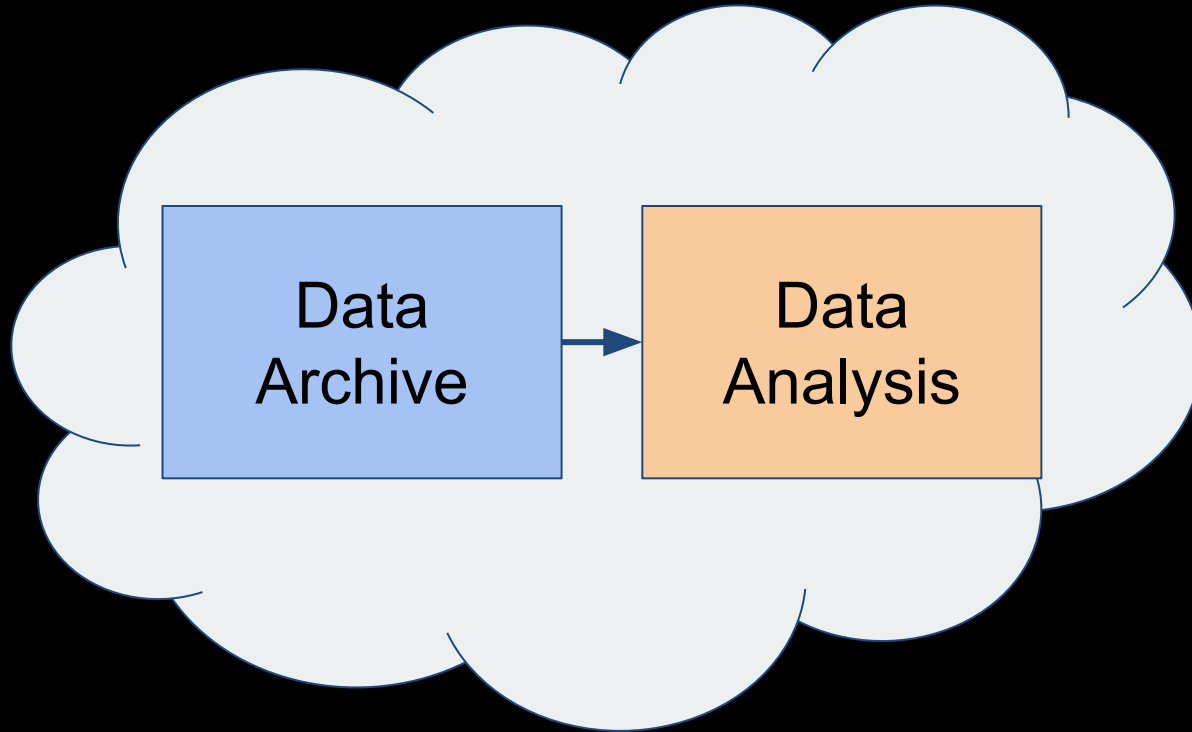






How do we remove the pain points to make  
Level 2 Swath data both Analysis-Ready and  
Analytics-Optimized?

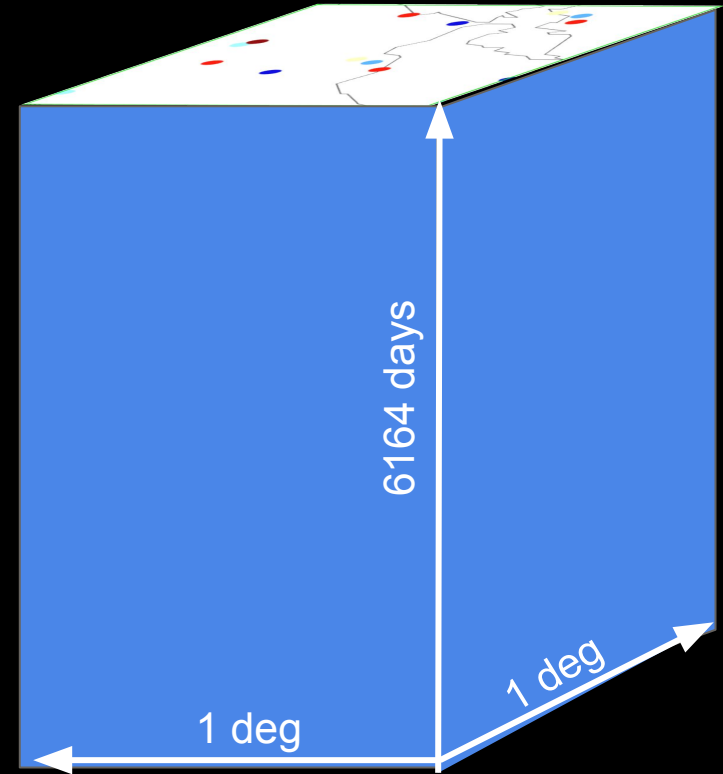
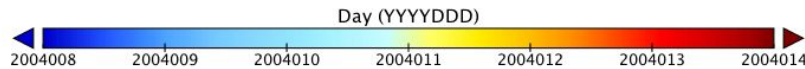
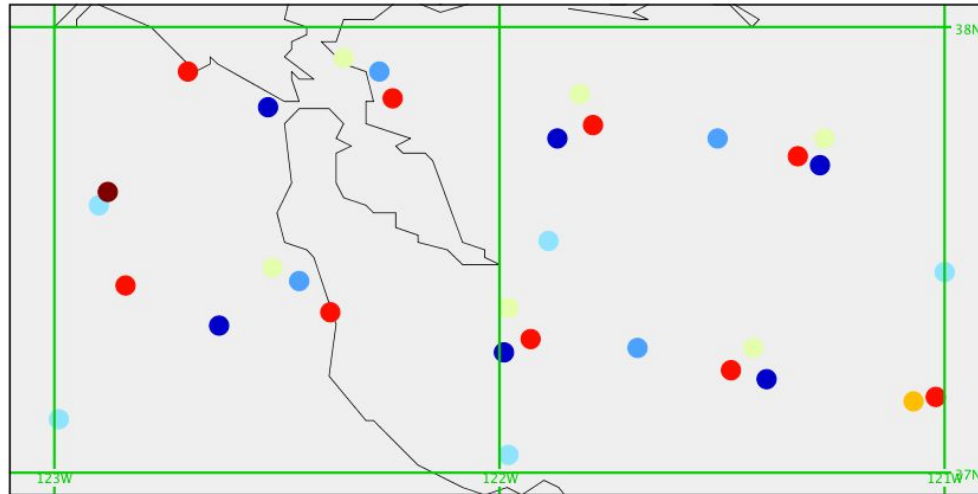
Step 1: Cloudify data and analysis to analyze in-place, without moving the data.



# Step 2: Reorganize Level 2 swath data to make spatial data access easier and faster



Pixel Center Locations for One Week of AIRS Observations  
(Nighttime)



# Columnar bins of L2 data: lat\_lon



```
$ ls -x TSurfAir.D
```

```
+24_-065    +24_-066    +24_-067    +24_-068    +24_-069
+24_-070    +24_-071    +24_-072    +24_-073    +24_-074
+24_-075    +24_-076    +24_-077    +24_-078    +24_-079
+24_-080    +24_-081    +24_-082    +24_-083    +24_-084
+24_-085    +24_-086    +24_-087    +24_-088    +24_-089
```

```
...
```

## Step 3: Use Python Application Program Interfaces to simplify spatial processing

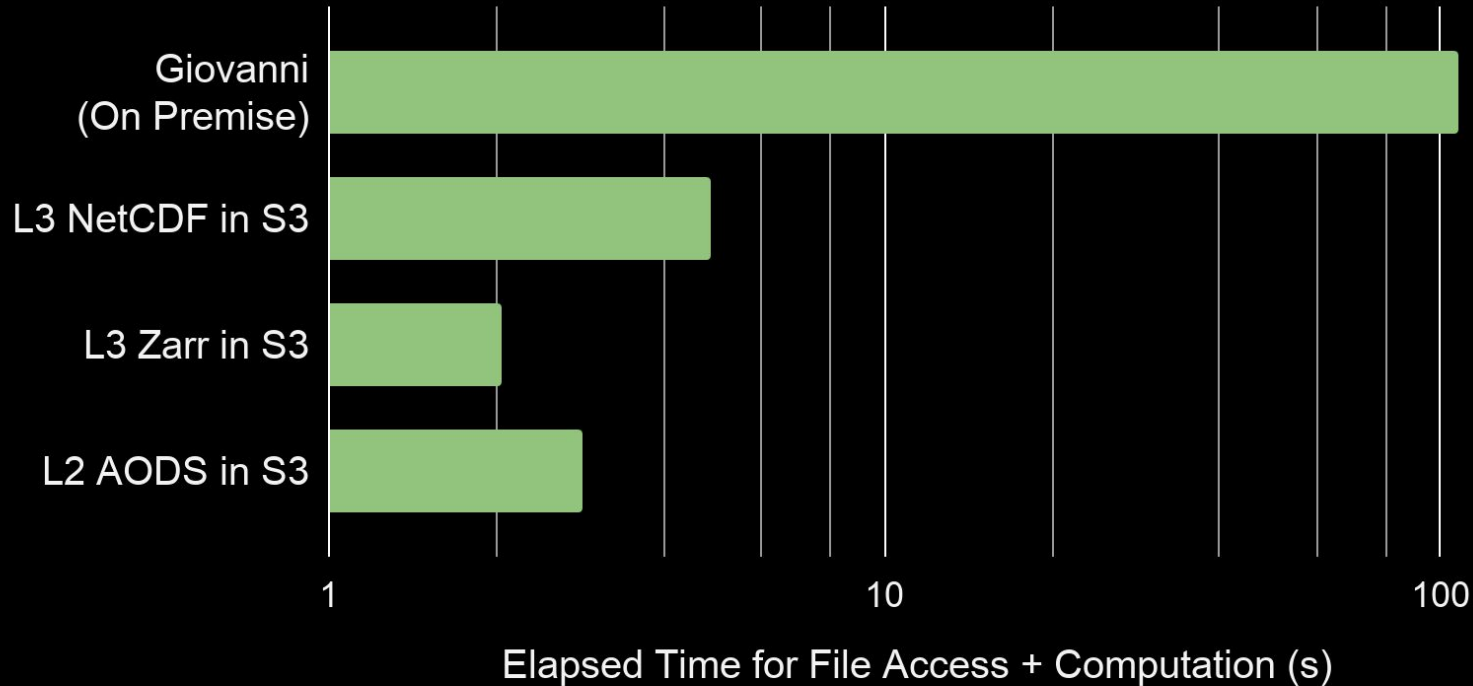


```
# Time series of data closest to a given point
ds['xdist'] = (ds['Longitude'] - my_lon)* \
              m.cos(my_lat * m.pi/180.)
ds['ydist'] = ds['Latitude']-my_lat
ds['sq_dist'] = xr.ufuncs.square(ds['ydist']) \
               + xr.ufuncs.square(ds['xdist'])
df = ds.to_dataframe()
closest_idx = df.groupby("dataday")['sq_dist'].idxmin()
closest = df.loc[closest_idx, ['dataday', 'TSurfAir']]
...
```

# How well does this L2 AODS perform?



Time series of data closest to a given point



NetCDF = Network Common Data Form    S3 = Simple Storage Service



# How do we make swath (L2) data easier and more performant to use?

1. Move data and analysis to the Cloud
2. Reorganize data for faster spatial access
3. Encapsulate geolocation irregularity with a Python API for geospatial access and computing
4. *Examples.* Lots and lots of examples.