



EOSDIS

NASA'S EARTH OBSERVING SYSTEM
DATA AND INFORMATION SYSTEM

Relevancy Ranking of Satellite Dataset Search Results

WGISS 2017

Christopher Lynnes (NASA EOSDIS)

Patrick Quinn (Element 84)

James Norton (Element 84)

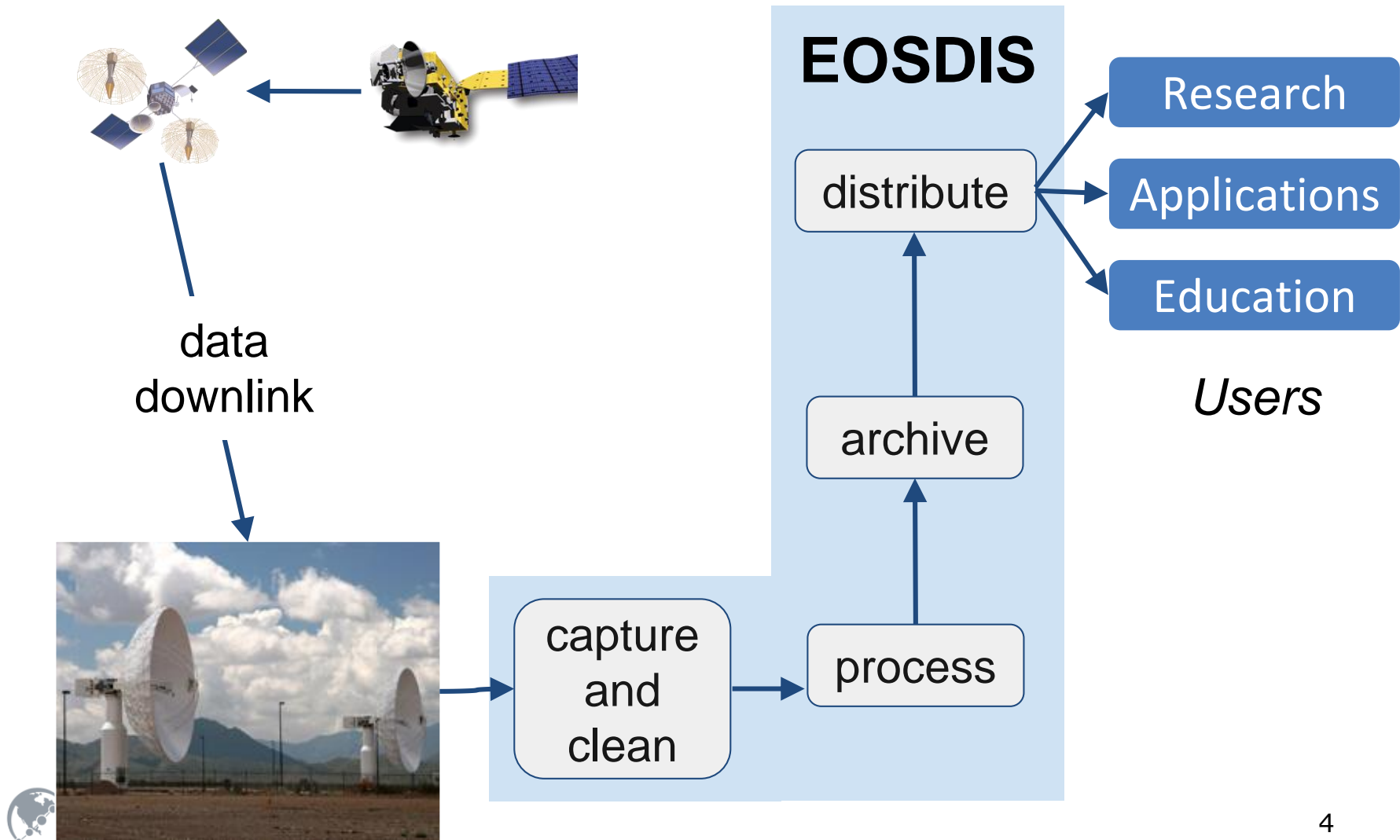
The Variety problem in Big Data from Satellites

Variety = Choice

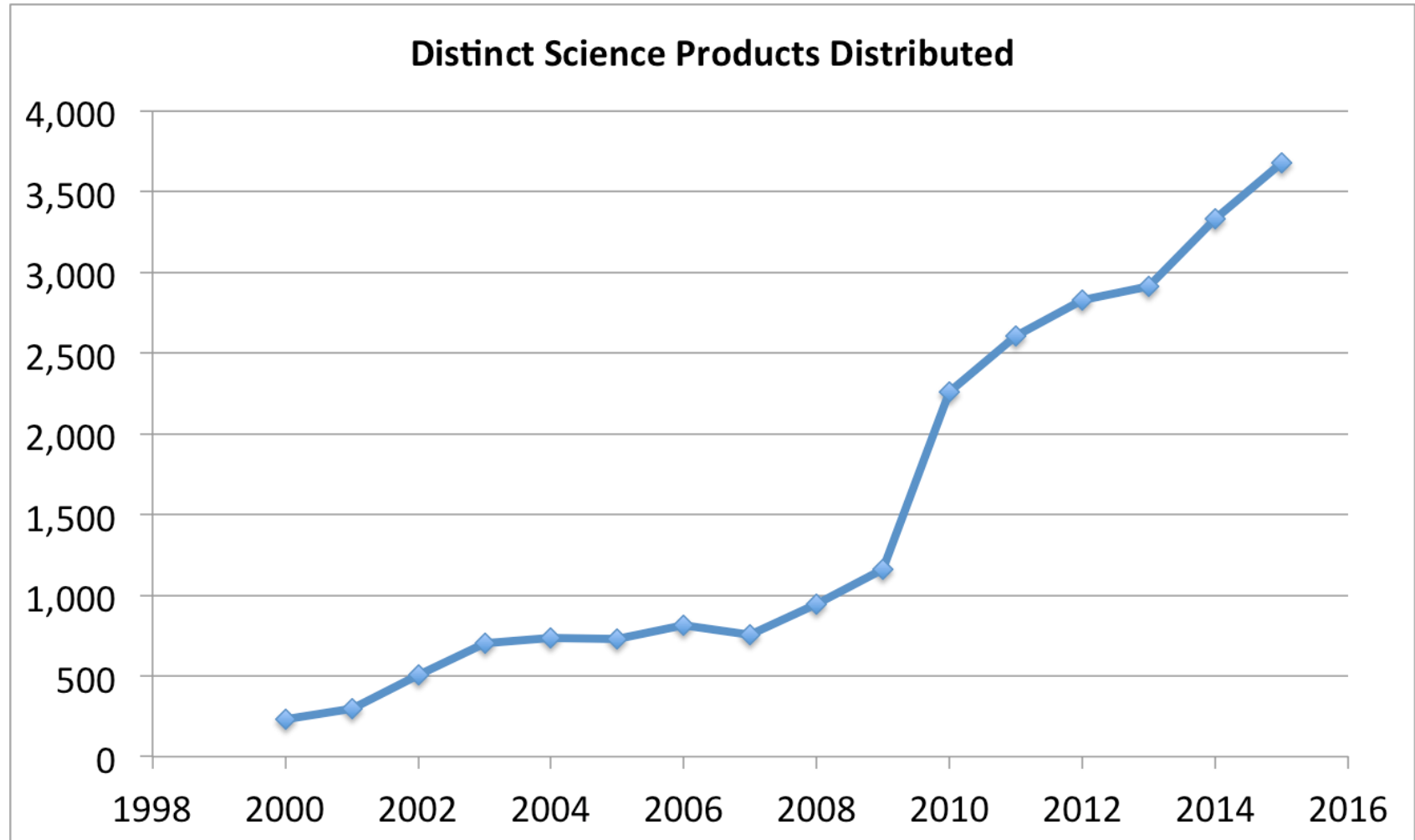
Choice = Good

(Right?)

The Earth Observing System Data and Information System (EOSDIS)



The Variety problem in Big Earth Data from Satellites





Earthdata Search




Discover Earth Science Data

[Take a Tour](#)

Search NASA Earth Science data by keyword and filter by  time or  space.



 [Browse All Data](#)

 See featured collections or use categories to narrow your results.

Too many datasets to sift manually




The screenshot shows the NASA EarthData Search interface. At the top left is the NASA logo and the text 'EARTHDATA Search'. A search bar contains the word 'Ozone'. To the right of the search bar is a 'Temporal' filter button. Below the search bar is a 'Browse Collections' button. A large green box highlights the text '1084 Matching Collections'. Below this, there is a message: 'Add collections to your project to compare and retrieve their data.' with a 'Learn More' button. A 'Search Time: 0.9s' indicator and a 'Report a metadata problem' button are also visible. The 'Recent and Featured' section displays a dataset card for 'BUV/Nimbus-4 Ozone (O3) Profile and Total Column Ozone 1 Month Zonal Mean L3 Global 5.0 degree Latitude Zones V1 (BUVN04L3zm) at GES DISC'. The card includes a 'No image available' placeholder, the dataset name, the source 'NASA/GSFC/SED/ESD/GCDC/GESDISC', the time range '1970-04-10 to 1976-05-01 | 1 Granule', and information and add buttons.

1084 Matching Collections

Add collections to your project to compare and retrieve their data. [Learn More](#)

Search Time: 0.9s [Report a metadata problem](#)

Recent and Featured



BUV/Nimbus-4 Ozone (O3) Profile and Total Column Ozone 1 Month Zonal Mean L3 Global 5.0 degree Latitude Zones V1 (BUVN04L3zm) at GES DISC

BUVN04L3zm v1 - NASA/GSFC/SED/ESD/GCDC/GESDISC

1970-04-10 to 1976-05-01 | 1 Granule

[i](#) [+](#)

Where does Variety come from?



Instruments

Fundamental differences: sounders, limb sounders, imagers...
Incremental evolution in instrument design

Satellites

“Same” instrument on different satellites

Processing Level

Calibrated -> Swath -> Grid -> Model

Processing Algorithm

Different basic principles
Incremental evolution in algorithm development

Temporal Resolution

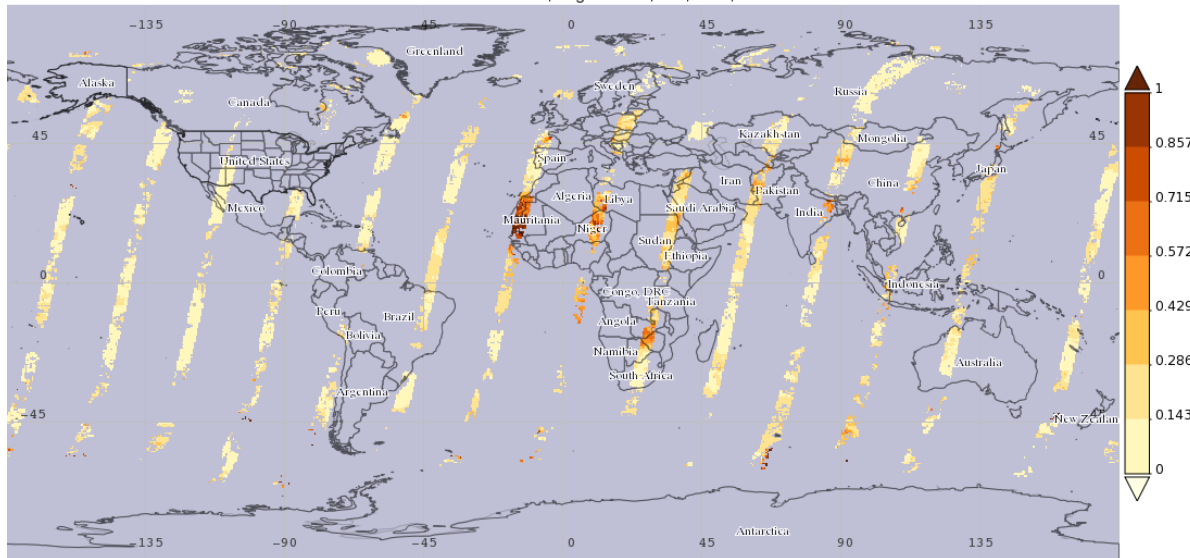
daily, 5-day, 8-day, monthly, yearly

Spatial Resolution...

Example: Time Aggregation



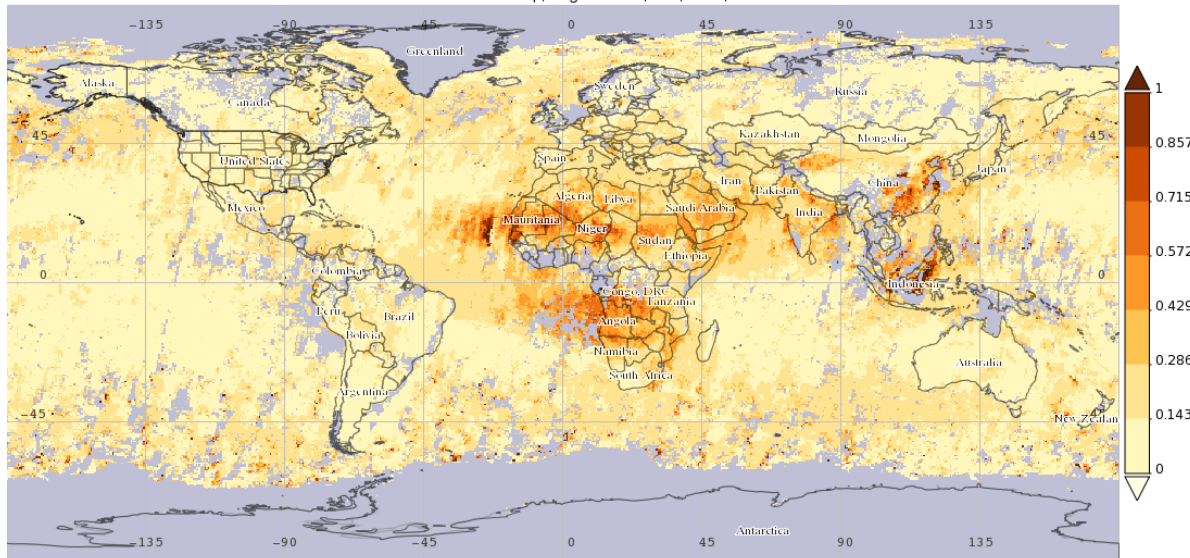
Time Averaged Map of Aerosol Optical Depth 555 nm daily 0.5 deg. [MISR MIL3DAE v4]
over 2009-09-21, Region 180W, 90S, 180E, 90N



Aerosol Optical Depth at 555 nm from Multi-angle Imaging Spectro-Radiometer

Daily

Time Averaged Map of Aerosol Optical Depth 555 nm monthly 0.5 deg. [MISR MIL3MAE v4]
over 2009-Sep, Region 180W, 90S, 180E, 90N



Monthly

- Selected date range was 2009-09-21 - 2009-09-21. Title reflects the date range of the granules that went into making this result.

What to do?



Emulate the best search engines: return the most relevant results at the top of the list

A la Wikipedia

“how well a retrieved document or set of documents meets the information need of the user”

HOW?






Relevancy Ranking Heuristics

Heuristic = “rule of thumb”

Basis is 20+ years of serving satellite data
to researchers

The Content Heuristic*

Got ozone?

Datasets Catalogs Bookmarks		
Name	Long Name	Type
▼  OMI-Aura_L3-OMTO3e_20...	OMI-Aura_L3-OMTO3e_20...	Remo...
 ColumnAmountO3	Best Total Ozone Solution	Geo2D
 lat	lat	1D
 lon	lon	1D
 RadiativeCloudFraction	Radiative Cloud Fraction = ...	Geo2D

“New-and-improved” Heuristics

New-and-Improved Processing Version

The screenshot displays two data product entries in a dark-themed interface. Each entry includes a placeholder for an image (a globe icon with the text 'No image available'), a title, a version identifier, a source, a date range, and a granule count. The version identifiers 'V004 (ML2O3)' and 'V003 (ML2O3)' are circled in green. Below each entry are two buttons: an information icon and a plus sign.

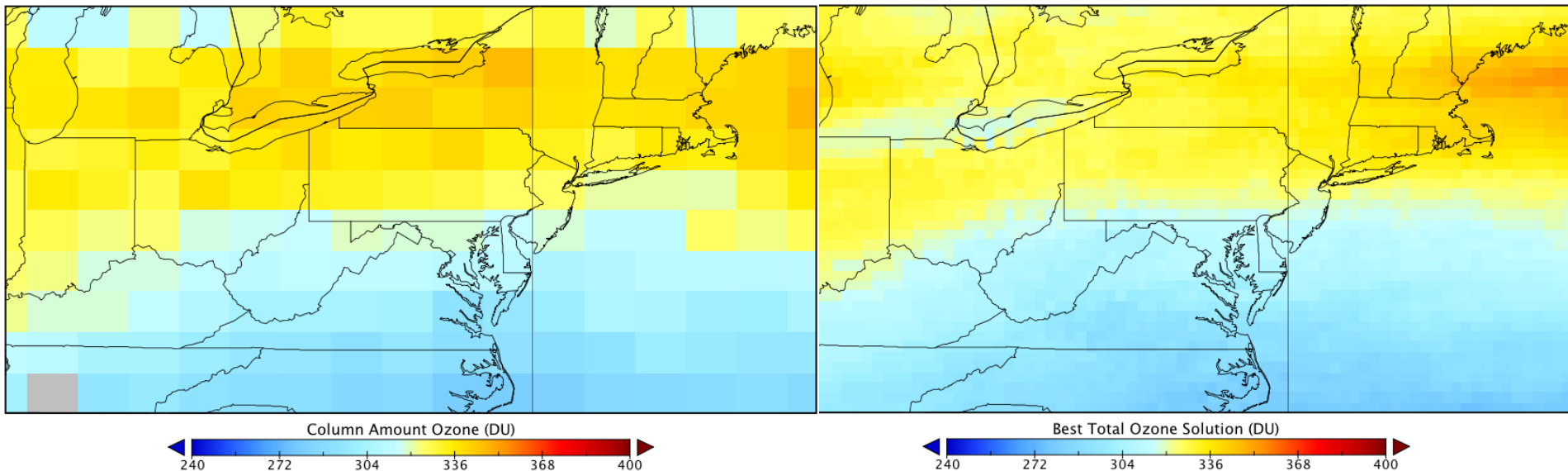
Product Name	Version	Source	Time Period	Granules
MLS/Aura Level 2 Ozone (O3) Mixing Ratio V004 (ML2O3) at GES DISC	V004 (ML2O3)	NASA/GSFC/SED/ESD/GCDC/GESDISC	2004-08-08 ongoing	4280
MLS/Aura Level 2 Ozone (O3) Mixing Ratio V003 (ML2O3) at GES DISC	V003 (ML2O3)	NASA/GSFC/SED/ESD/GCDC/GESDISC	2004-08-08 to 2015-06-30	3935

New processing version is also more likely to be up to date

MLS/Aura Level 2 Ozone (O3) Mixing Ratio V004 (ML2O3) at GES DISC
ML2O3 v004 - NASA/GSFC/SED/ESD/GCDC/GESDISC
2004-08-08 to ongoing | 4280 Granules

MLS/Aura Level 2 Ozone (O3) Mixing Ratio V003 (ML2O3) at GES DISC
ML2O3 v003 - NASA/GSFC/SED/ESD/GCDC/GESDISC
2004-08-08 to 2015-06-30 | 3935 Granules

Newer instrument is usually better than previous instruments



Total Ozone Mapping Spectrometer

Ozone Monitoring Instrument

Region of Interest Overlap

Time Range Heuristic

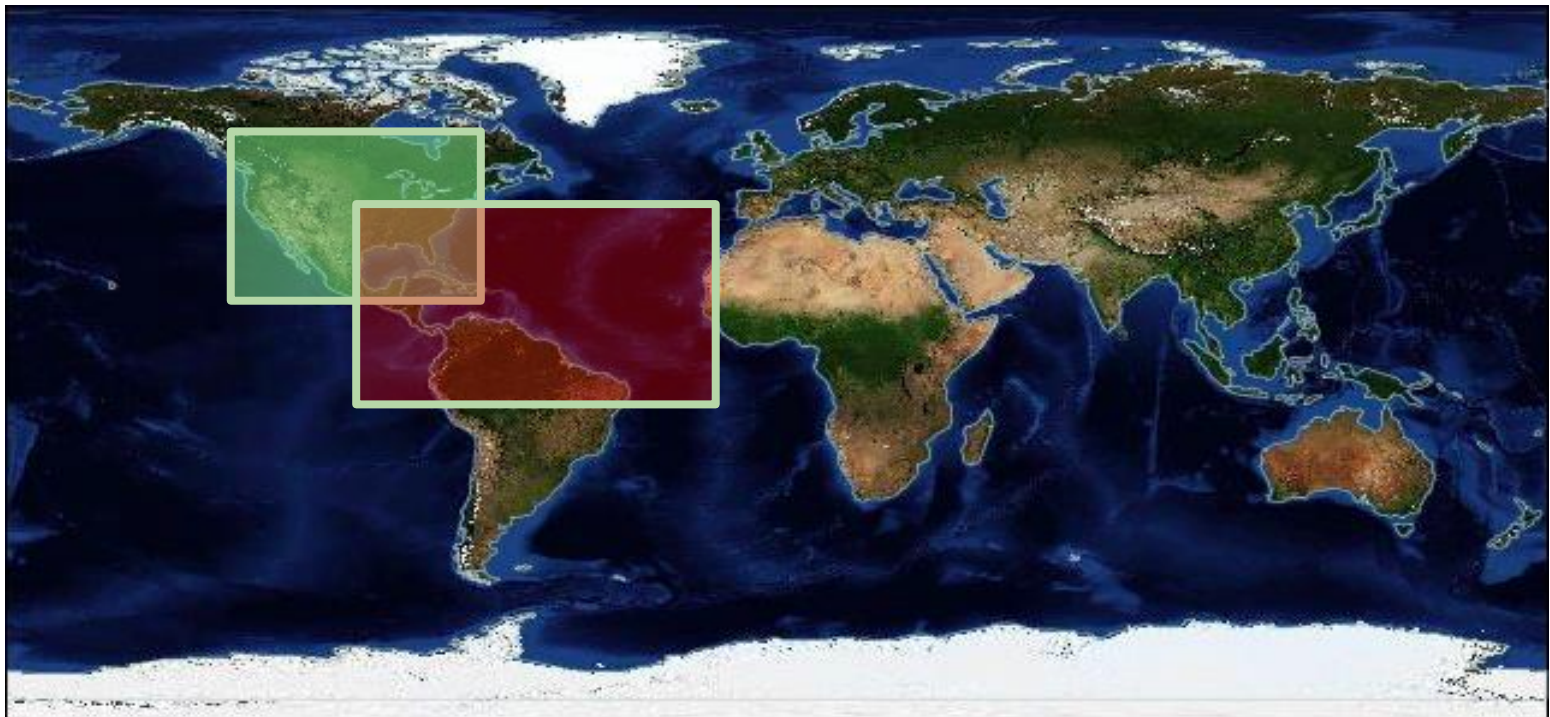
Datasets covering the user's full time range are better than those covering just part of it

	2005	2006	2007	2008	2009	2010	
<i>Time range of interest</i>							
TOMS-Earth Probe							Meh.
Ozone Monitoring Inst.							Yeah!

Spatial Heuristic

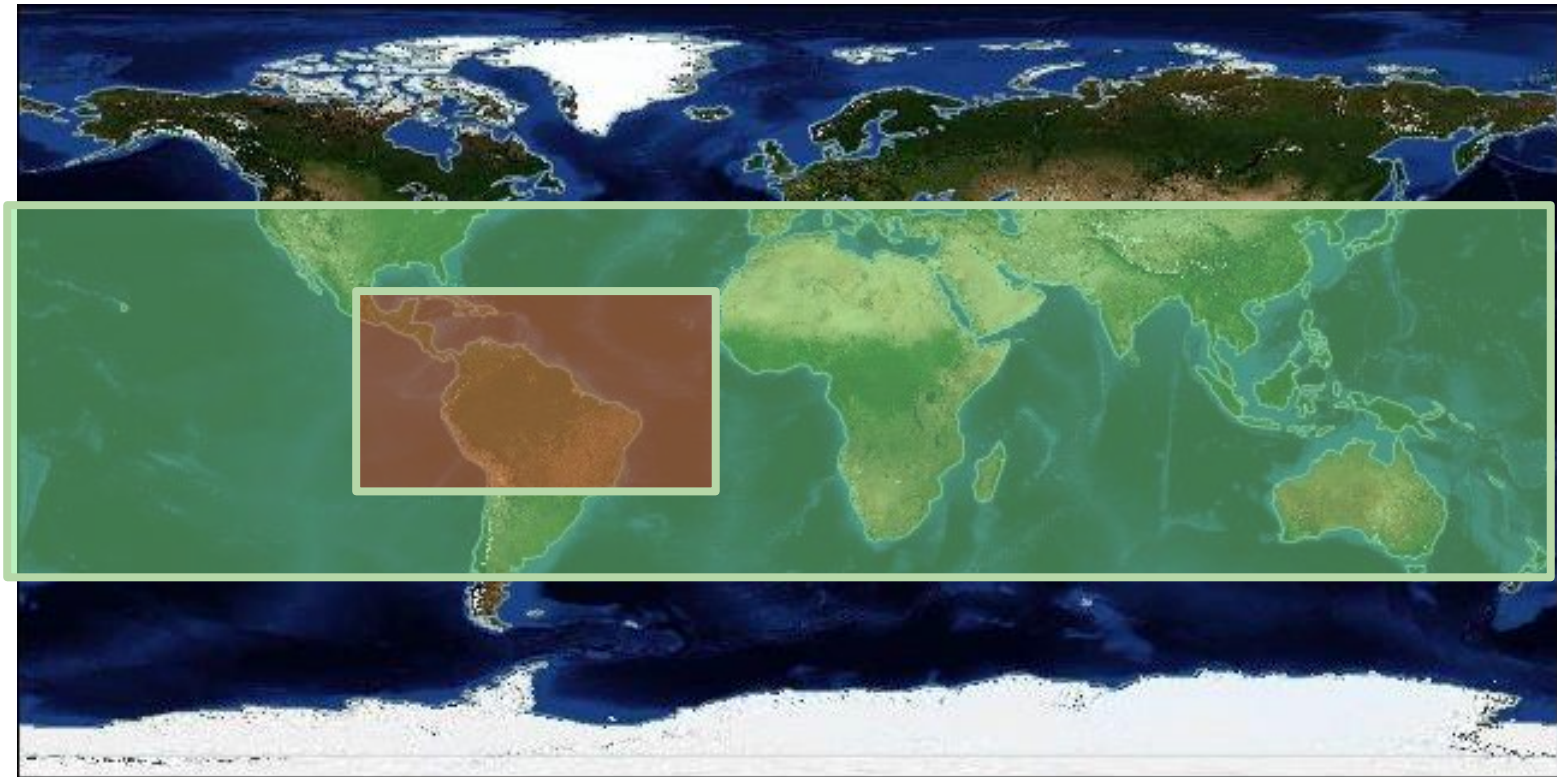
Data covering the user's full area are better than those covering just part of it.

This is not as good as...



Spatial Heuristic

This...



User-centric Heuristics

Community Usage Heuristic

The dataset most often used by the community is more likely to be useful

Data Product	Users**
Aqua AIRS Level 3 Daily Standard Physical Retrieval (AIRS only)*	164
Aqua AIRS Level 3 Daily Standard Physical Retrieval (AIRS+AMSU)*	714

*Version 6

** Jan 1, 2016 - June 20, 2016

User Intent Heuristics

User type or intent*	The most relevant datasets are...
Applications users	High spatial resolution, near-real-time
Students	Easier to use data <i>e.g., L3 grids in netCDF</i>
Climate Modeler	Datasets on Climate Model Grid