Rensselaer

why not change the world?®

Data Quality, Uncertainty and Bias

Ahmed Eleish

Data Science – ITWS/CSCI/ERTH-4350/6350

October 30th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Defining quality
- How fast it gets complicated
- A couple of examples
- Advanced approaches

- How are the projects going?

Accuracy & Precision

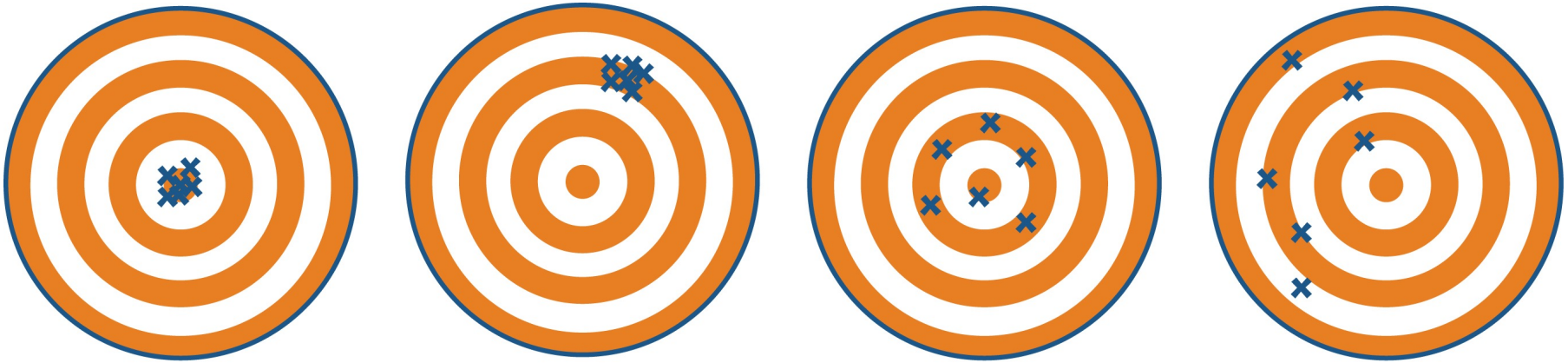
- **Accuracy:**

  Accuracy is how close a measured value is to the **actual (true) value**.

- **Precision:**

  Precision is how close the measured values are **to each other**.

Accurate vs. Precise



High Accuracy
High Precision

Low Accuracy
High Precision

High Accuracy
Low Precision

Low Accuracy
Low Precision

http://climatica.org.uk/climate-science-information/uncertainty

Example: Accuracy Vs Precision

If you are playing football (soccer lol) and you always hit the right goal post instead of scoring, then you are **not** accurate, but you **are** precise!

Bias

- **Bias (don't let precision fool you!)**
- When we measure something several times and all values are close, they **may** all be wrong if there is a "**Bias**"
- **Bias** is a systematic (built-in) error which makes all measurements wrong by a certain amount.

Bias:

Examples of Bias:

- The scales read "1 kg" when there is nothing on them
- You always measure your height wearing shoes with thick soles.
- A stopwatch that takes half a second to stop when the button is pressed

- In each case **all** measurements are wrong by the same amount. That is bias.

Degree of Accuracy

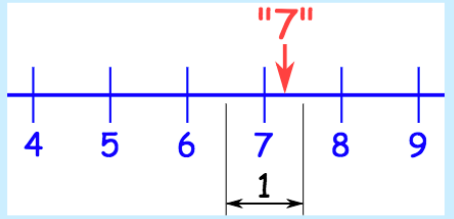
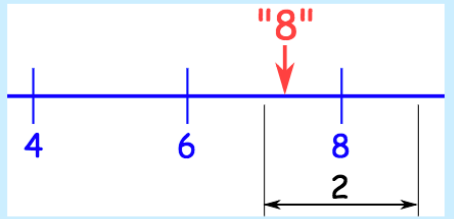
- Accuracy depends on the instrument we are measuring with. But as a general rule:

The degree of accuracy is **half a unit** each side of the unit of measure

Degree of Accuracy

Notice that the arrow points to the same spot, but the measured values are different!

Read more at Errors in Measurement.

When an instrument measures in "1"s any value between **6½** and **7½** is measured as **"7"**	
When an instrument measures in "2"s any value between **7** and **9** is measured as **"8"**	

https://www.mathsisfun.com/measure/error-measurement.html

Modes of Data Collection

- Remember the modes of data collection?

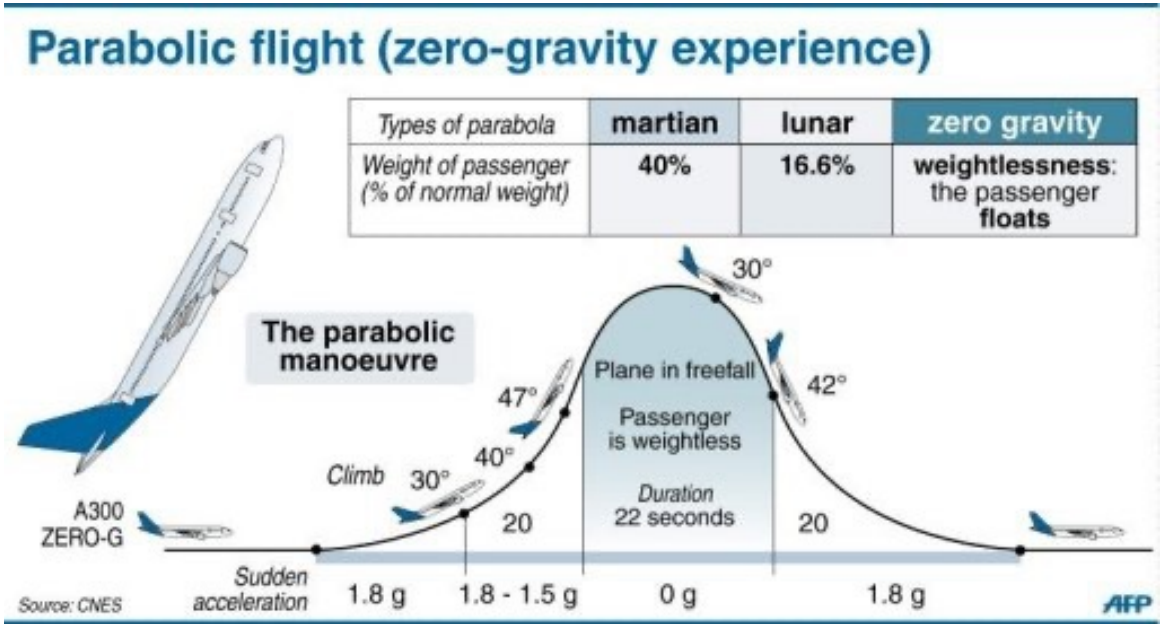
- Observation
- Measurement
- Generation

Microgravity Experiment

Generating Pool Boiling Data in Zero Gravity

- Generating pool boiling data in Zero Gravity (Micro-gravity)
- Zero Gravity = Outer space
- What is Micro-gravity?
- Can you boil water in space using conventional boiling techniques?
- How can we generate (or do simulations) of boiling data in microgravity condition/environment?

Simulating Zero Gravity



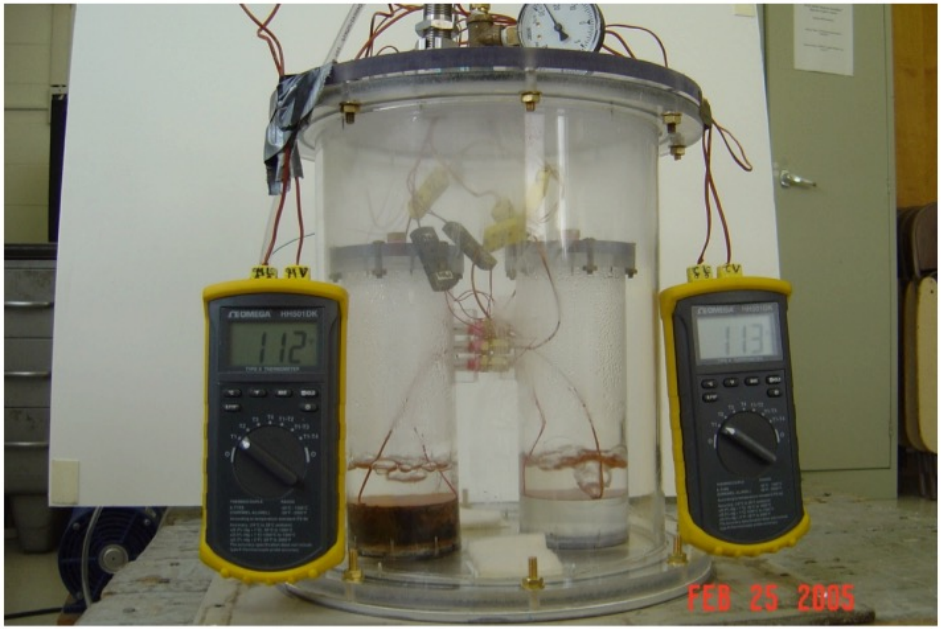


Figure 6: Tanks with Omega HH5010K 4-Channel Thermocouple Readers

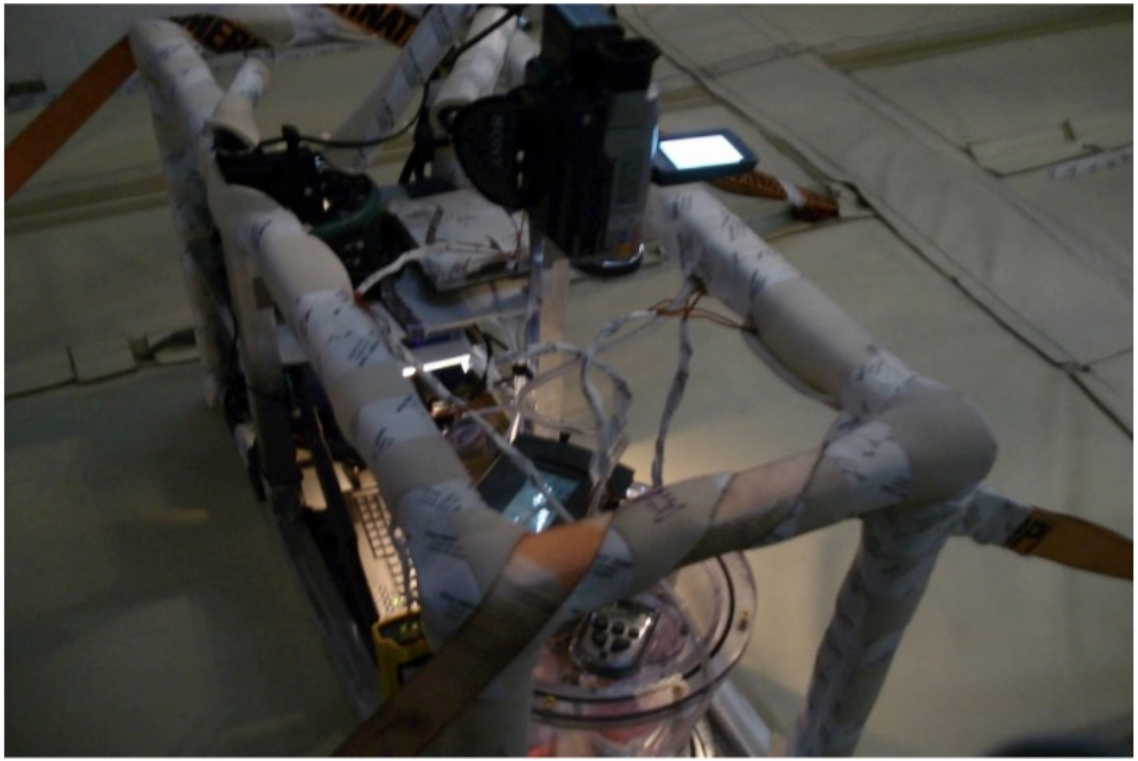


Figure 11: Final Experiment Setup Before the Flight

Microgravity Environment

- Microgravity Experience inside ZeroG flight
- Conducting the research experiment in an artificially generated microgravity environment.
- https://www.youtube.com/watch?v=1wNK6oz9Pck

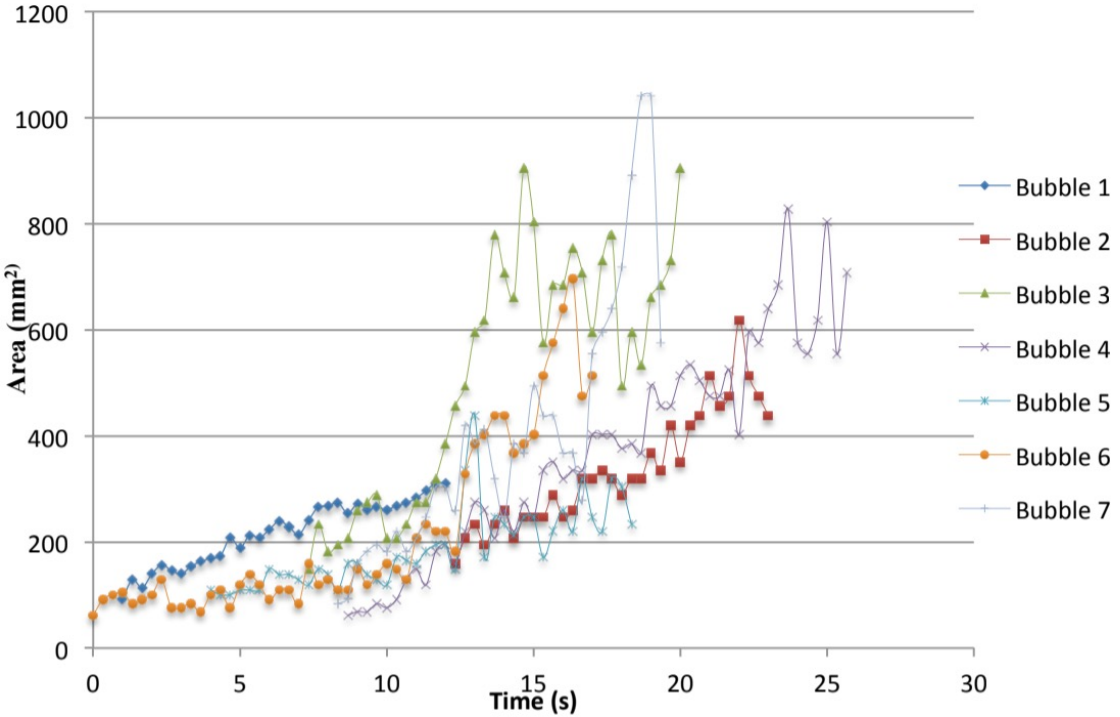


Figure 25: Areas of Bubbles with Respect to Time (Magnet Tank); Parabola 35

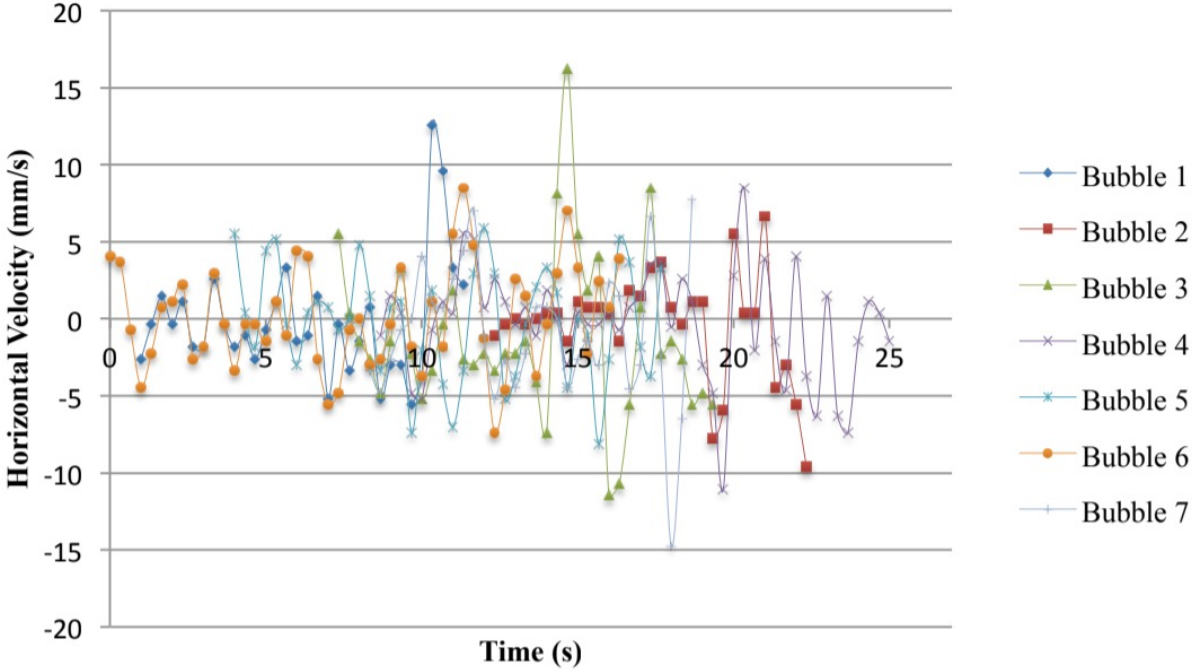


Figure 30: Visual Representation of the Horizontal Velocities of the Bubbles that are Located in Magnet Tank; Parabola 35

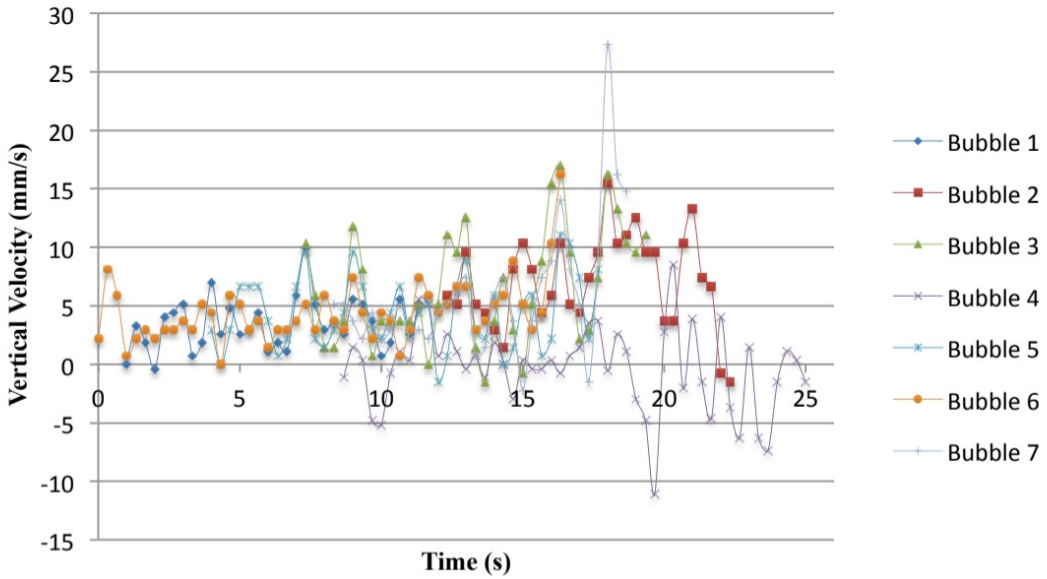


Figure 31: Vertical Velocity (Magnet Tank); Parabola 35.

Thus results of the study indicate:

- Under the magnetic field influence, bubbles which were generated by pool boiling under the microgravity conditions are more circular in their cross sectional shape, and are smaller than in the non-magnet tank.
- On the other hand, bubbles without the magnetic field influence tend to have larger horizontal radius than the vertical radius, and are larger than in the magnet tank.
- Under the microgravity conditions, the boiling process inside the two tanks take place suddenly in an explosive manner within a short period of time as explained by the velocity graphs.

Definitions – for an atmospheric scientist

- **Quality**
  - Is in the eyes of the beholder – worst case scenario… or a good challenge
- **Uncertainty**
  - has aspects of accuracy (how accurately the real world situation is assessed, it also includes bias) and precision (down to how many digits)
- **Bias** has two aspects:
  - Systematic error resulting in the distortion of measurement data caused by prejudice or faulty measurement technique
  - A vested interest, or strongly held paradigm or condition that may skew the results of sampling, measuring, or reporting the findings of a quality assessment:
    - Psychological: for example, when data providers audit their own data, they usually have a bias to overstate its quality.
    - Sampling: Sampling procedures that result in a sample that is not truly representative of the population sampled. (Larry English)

Data quality & fitness-for-use

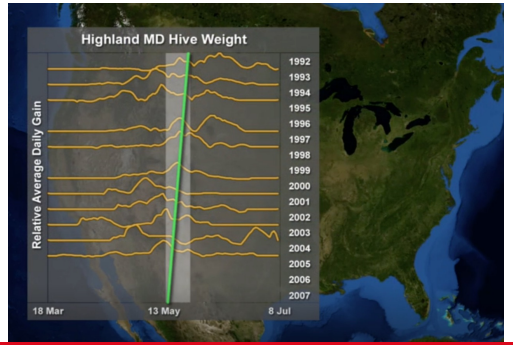
- Short introduction to biodiversity data quality and fitness-for-use

https://vimeo.com/40443456

Modes of Data Collection

- Observation
- Measurement

Comparing bee data to satellite imagery data



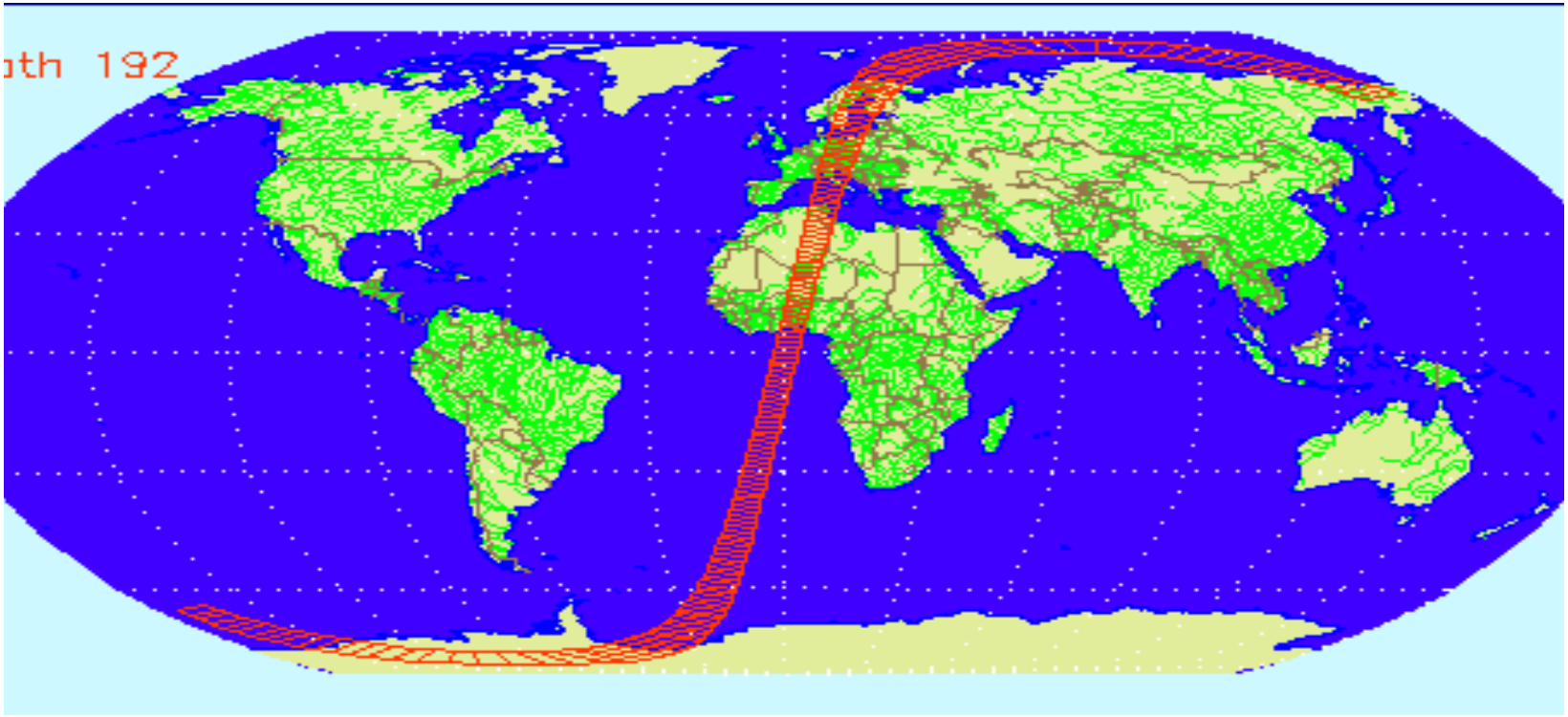
Comparing Data:

Findings: Comparing Bee Data to Satellite Imagery Data

- https://climate.nasa.gov/climate_resources/41/video-sting-of-climate-change/

Data quality needs: fitness for purpose

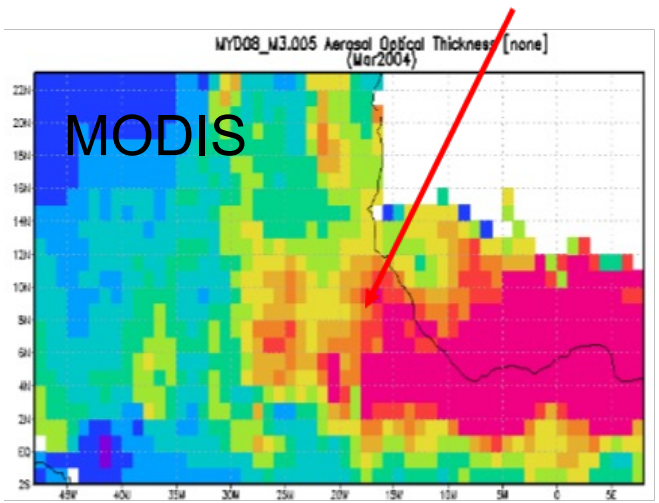
- **Measuring Climate Change:**
  - *Model validation*: **gridded contiguous data with uncertainties**
  - *Long-term time series*: **bias assessment** is the must , especially sensor degradation, orbit and spatial sampling change
- **Studying phenomena using multi-sensor data:**
  - **Cross-sensor bias** is needed
- **Realizing Societal Benefits through Applications:**
  - *Near-Real Time for transport/event monitoring* - in some cases, **coverage and timeliness** might be more important that accuracy
  - *Pollution monitoring* (e.g., air quality exceedance levels) – **accuracy**
- **Educational** (users generally not well-versed in the intricacies of quality; just taking all the data as usable can impair educational lessons) – **only the best products**

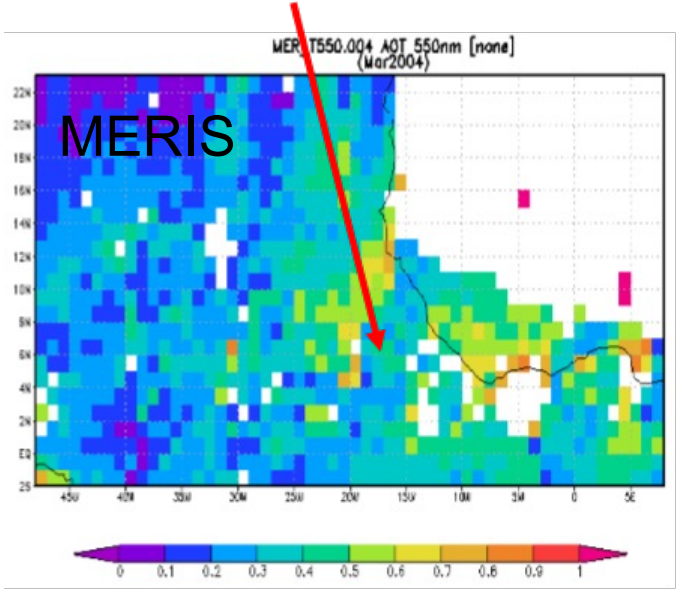

ıth 192

MODIS vs. MERIS

Same parameter

Same space & time





Different results – why?

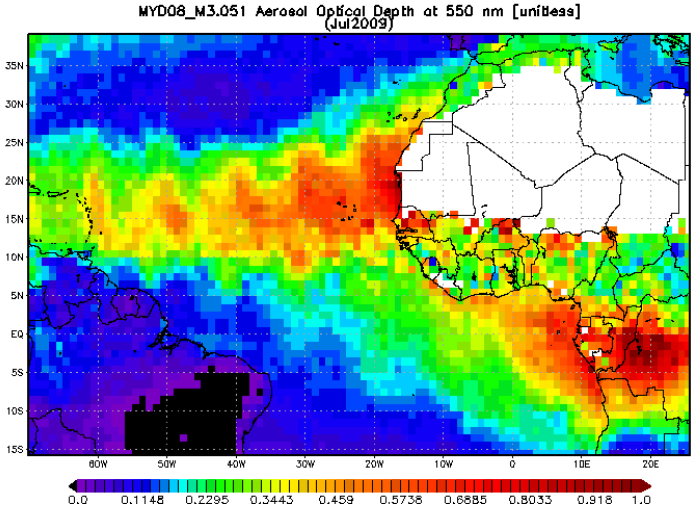
A threshold used in MERIS processing effectively excludes high aerosol values. *Note: MERIS was designed primarily as an ocean-color instrument, so aerosols are "obstacles" not signal.*

Why so difficult?

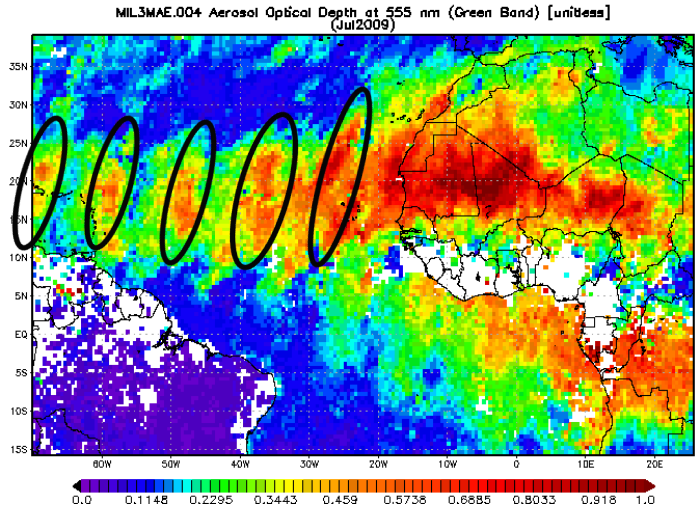
- Quality is **perceived differently** by data providers and data recipients.
- There are **many different qualitative and quantitative aspects** of quality.
- **Methodologies** for dealing with data qualities are still **emerging.**
- Very little exists for **remote sensing data quality.**
- Even the most comprehensive review - Batini's book (look it up) demonstrates that there are **no preferred methodologies** for solving many data quality issues.
- Little funding was allocated in the past to data quality as the priority was to build an instrument, launch a rocket, collect and process data, and publish a paper using just one set of data.
- Scientists/researchers handle quality differently.

Spatial and temporal sampling – how to quantify to make it useful for modelers?

MODIS Aqua AOD July 2009



MISR Terra AOD July 2009



- *Completeness*: MODIS dark target algorithm does not work for deserts
- *Representativeness*: monthly aggregation is not enough for MISR and even MODIS
- *Spatial sampling* patterns are different for MODIS Aqua and MISR Terra: "pulsating" areas over ocean are oriented differently due to different orbital direction during day-time measurement → *Cognitive bias*

More terminology

- ‘Even a slight difference in terminology can lead to significant differences between data from different sensors. It gives an IMPRESSION of data being of bad quality while in fact they measure different things. For example, MODIS and MISR definitions of the aerosol "fine mode" is different, so the **direct comparison** of fine modes **from MODIS and MISR does not always give good correlation.**’

- Ralph Kahn, MISR Aerosol Lead.

Quality Control vs. Quality Assessment

- Quality Control (QC) flags in the data (assigned by the algorithm) reflect “happiness” of the retrieval algorithm, e.g., all the necessary channels indeed had data, not too many clouds, the algorithm has converged to a solution, etc.

- Quality assessment is done by analyzing the data “after the fact” through validation, intercomparison with other measurements, self-consistency, etc. It is presented as bias and uncertainty. It is rather inconsistent and can be found in papers, validation reports all over the place.

Intercomparison of data from multiple sensors

Data from multiple sources to be used together:

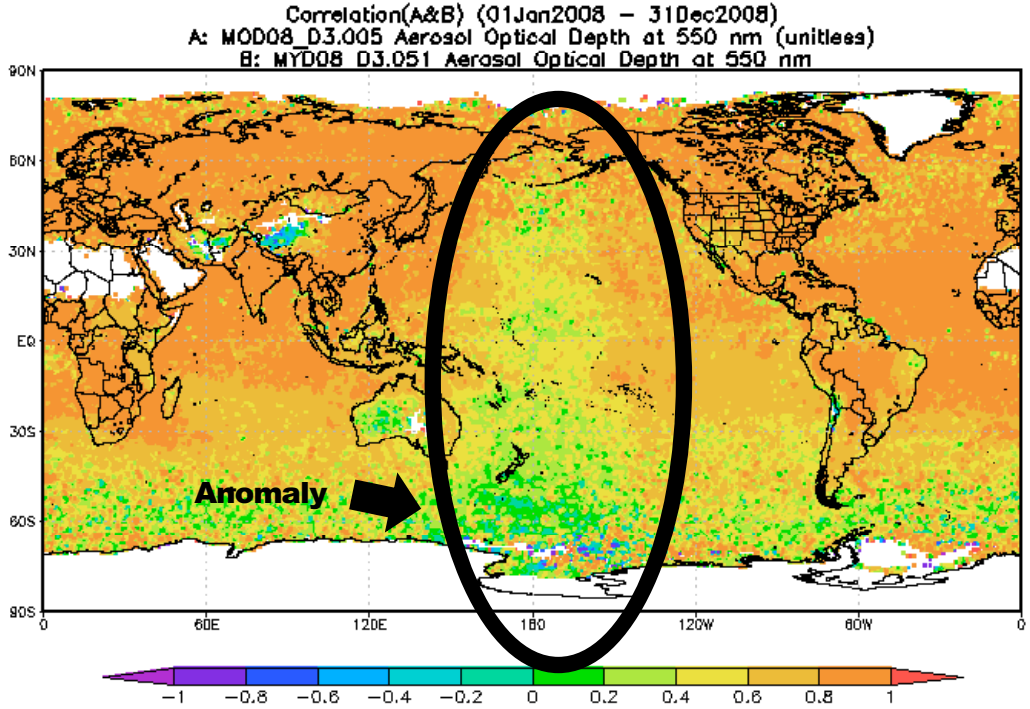
- Current sensors/missions: MODIS, MISR, GOES, OMI

Harmonization needs:

- It is not sufficient just to have the data from different sensors and their provenances in one place
- Before comparing and fusing data, things need to be harmonized:
  - Metadata: terminology, standard fields, units, scale
  - Data: format, grid, spatial and temporal resolution, wavelength, etc.
  - Provenance: source, assumptions, algorithm, processing steps
  - **Quality: bias, uncertainty, fitness-for-purpose, validation**

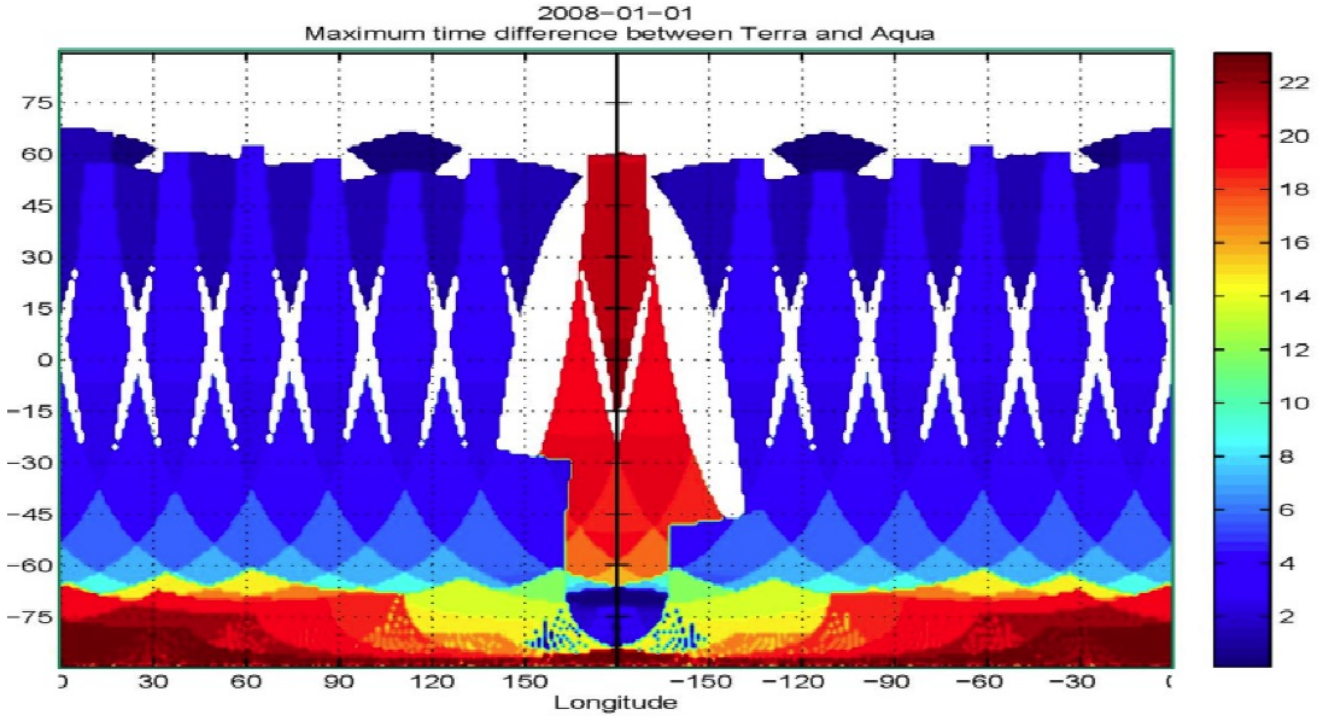
Dangers of easy data access without proper assessment of the joint data usage - *It is easy to use data incorrectly*

Anomaly Example: South Pacific Anomaly



MODIS Level 3 dataday definition leads to artifact in correlation

…is caused by an Overpass Time Difference



Different kinds of reported data quality

- **Pixel-level** Quality: algorithmic guess at usability of data point
    - Granule-level Quality: statistical roll-up of Pixel-level Quality
- **Product-level** Quality: how closely the data represent the actual geophysical state
- **Record-level** Quality: how consistent and reliable the data record is across generations of measurements

Different quality types are often erroneously assumed having the same meaning

Ensuring Data Quality at these different levels requires different focus and action

Three projects on data quality

- Multi-sensor Data Synergy Advisor - Product level
- Data Quality Screening Service - Pixel level
- Aerosol Statistics - Record level

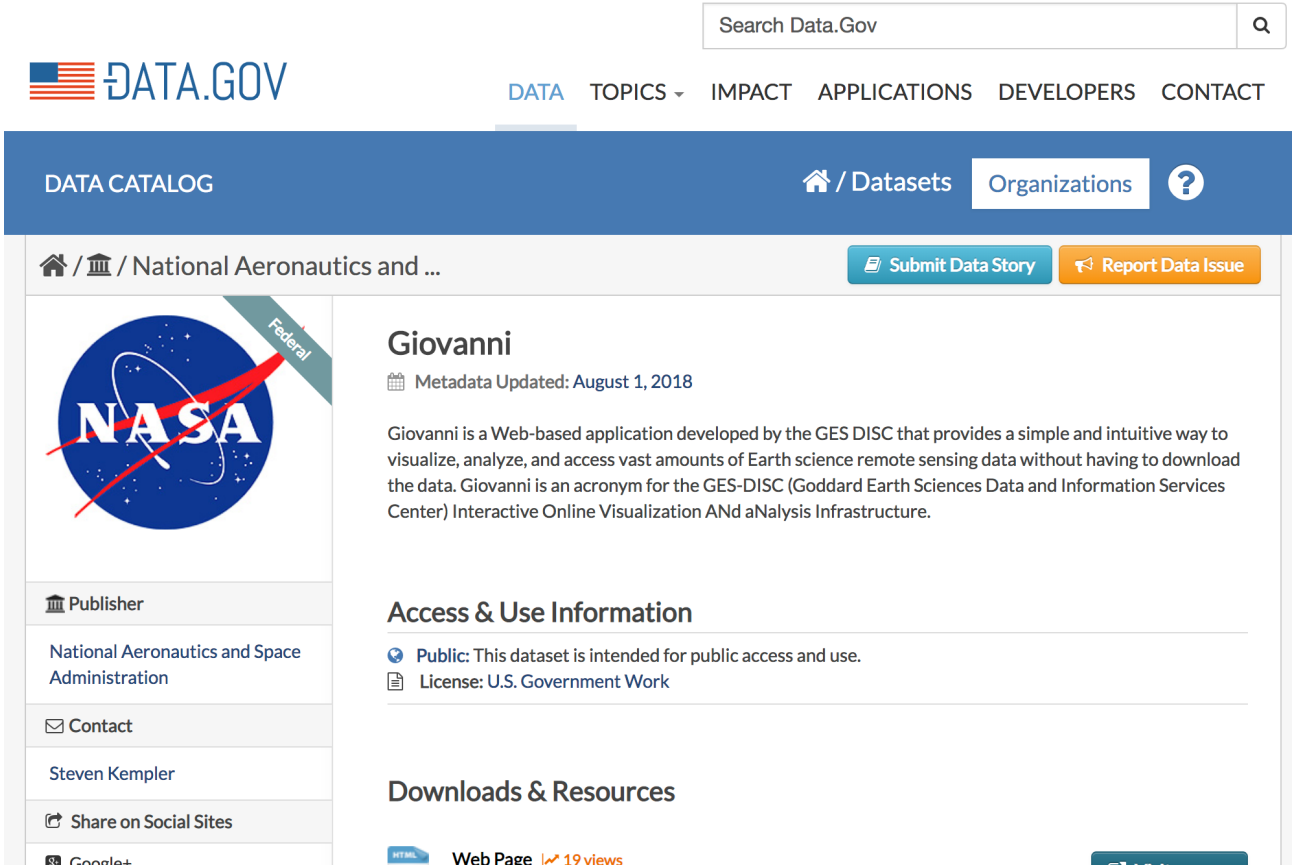
Multi-Sensor Data Synergy Advisor (MDSA)

- *Goal*: Provide science users with clear, cogent information on salient differences between data candidates for fusion, merging and intercomparison
    - Enable scientifically and statistically valid conclusions
- Develop MDSA on current missions:
    - NASA - Terra, Aqua, (maybe Aura)
- Define implications for future missions

How MDSA works?

MDSA is a service designed to characterize the differences between two datasets and advise a user (human or machine) on the advisability of combining them.

- Provides the Giovanni online analysis tool
- Describes parameter and products
- Documents steps leading to the final data product
- Enables better interpretation and utilization of parameter difference and correlation visualizations.
- Provides clear and cogent information on salient differences between data candidates for intercomparison and fusion.
- Provides information on data quality
- Provides advice on available options for further data processing and analysis.


Search Data.Gov
DATA.GOV
DATA
TOPICS
IMPACT
APPLICATIONS
DEVELOPERS
CONTACT
DATA CATALOG
/ Datasets
Organizations
/ National Aeronautics and ...
Submit Data Story
Report Data Issue
Federal
NASA
Giovanni
Metadata Updated: August 1, 2018
Giovanni is a Web-based application developed by the GES DISC that provides a simple and intuitive way to visualize, analyze, and access vast amounts of Earth science remote sensing data without having to download the data. Giovanni is an acronym for the GES-DISC (Goddard Earth Sciences Data and Information Services Center) Interactive Online Visualization ANd aNalysis Infrastructure.
Publisher
National Aeronautics and Space Administration
Contact
Steven Kempler
Share on Social Sites
Access & Use Information
Public: This dataset is intended for public access and use.
License: U.S. Government Work
Downloads & Resources
Web Page
19 views

Giovanni: An Easier Way to Visualize Earth Sci Data

Giovanni: An Easier Way to Visualize Earth Science Data **06.17.08**

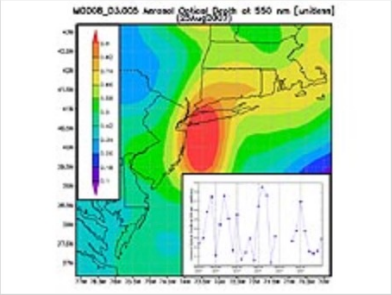
The world of data can be a confusing one. Different formats, hefty downloads and complicated plotting tools can bog down even the experienced researcher, let alone a teacher or student.

Giovanni is a tool that displays Earth science data from NASA satellites directly on the Internet, without the difficulties of traditional data acquisition and analysis methods. Giovanni is an acronym for the Goddard Earth Sciences Data and Information Services Center, or GES DISC, Interactive Online Visualization and Analysis Infrastructure.

With a few clicks, data from various instruments on NASA satellites can be displayed in a variety of formats, including area plots, time series, meridional averages, zonal averages and vertical profiles, among others. Animations and numerical outputs are also available.

Users can analyze phenomena ranging from the environment surrounding a Saharan dust storm to the impact Hurricane Katrina had on ocean surface chlorophyll concentrations. Single and multiple parameters can be plotted for specified ranges and time periods.

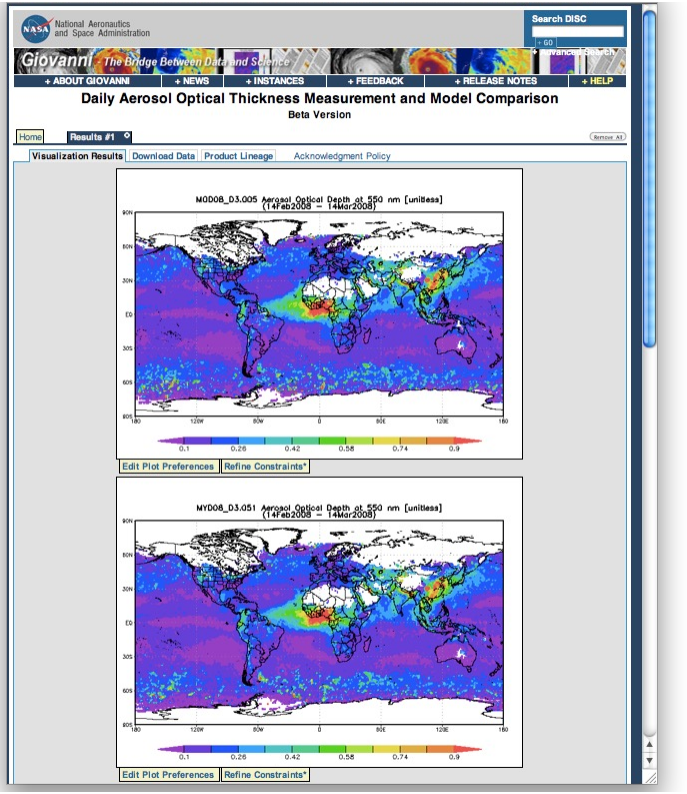
Data is accessed and displayed via a collection of interfaces, each one allowing the plotting of parameters from one or more satellite instruments. Rainfall, temperature, humidity and gas concentrations are just a few of the parameters provided.

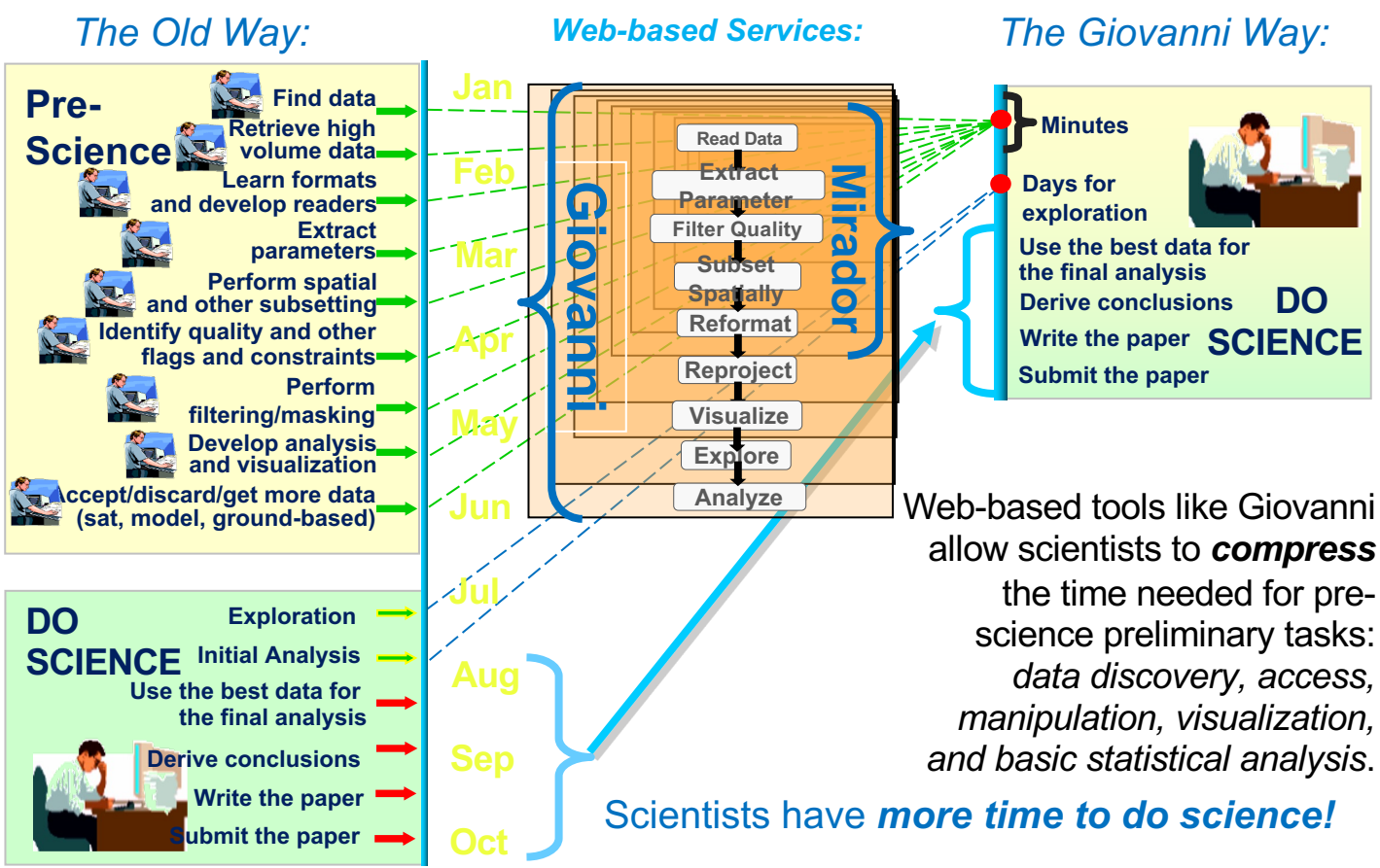


The Giovanni Web site allows users to analyze phenomena ranging from the environment surrounding a Saharan dust storm to the impact Hurricane Katrina had on ocean surface chlorophyll concentrations. Image Credit: NASA

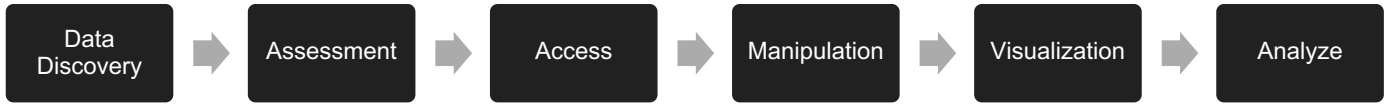
Giovanni Earth Science Data Visualization & Analysis Tool

- Developed and hosted by NASA/ Goddard Space Flight Center (GSFC)
- Multi-sensor and model data analysis and visualization online tool
- Supports dozens of visualization types
- Generate dataset comparisons
- ~1500 Parameters
- Used by modelers, researchers, policy makers, students, teachers, etc.



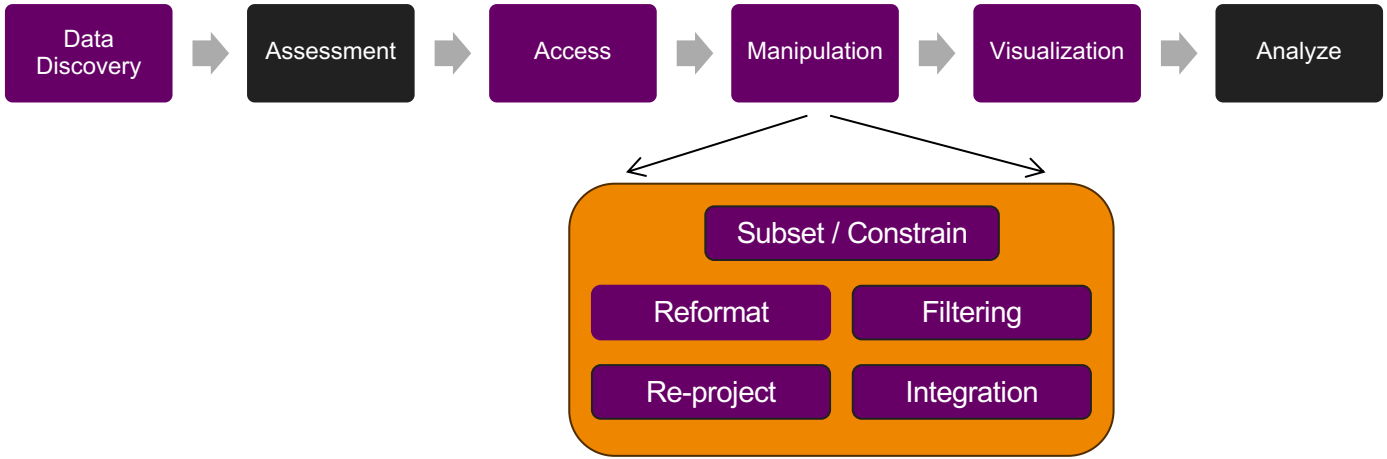

The Old Way:
Pre-Science
Find data
Retrieve high volume data
Learn formats and develop readers
Extract parameters
Perform spatial and other subsetting
Identify quality and other flags and constraints
Perform filtering/masking
Develop analysis and visualization
Accept/discard/get more data (sat, model, ground-based)
DO SCIENCE
Exploration
Initial Analysis
Use the best data for the final analysis
Derive conclusions
Write the paper
Submit the paper
Web-based Services:
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Giovanni
Mirador
Read Data
Extract Parameter
Filter Quality
Subset Spatially
Reformat
Reproject
Visualize
Explore
Analyze
The Giovanni Way:
Minutes
Days for exploration
Use the best data for the final analysis
Derive conclusions
Write the paper
Submit the paper
DO SCIENCE
Web-based tools like Giovanni allow scientists to compress the time needed for pre-science preliminary tasks: data discovery, access, manipulation, visualization, and basic statistical analysis.
Scientists have more time to do science!

Data Usage Workflow

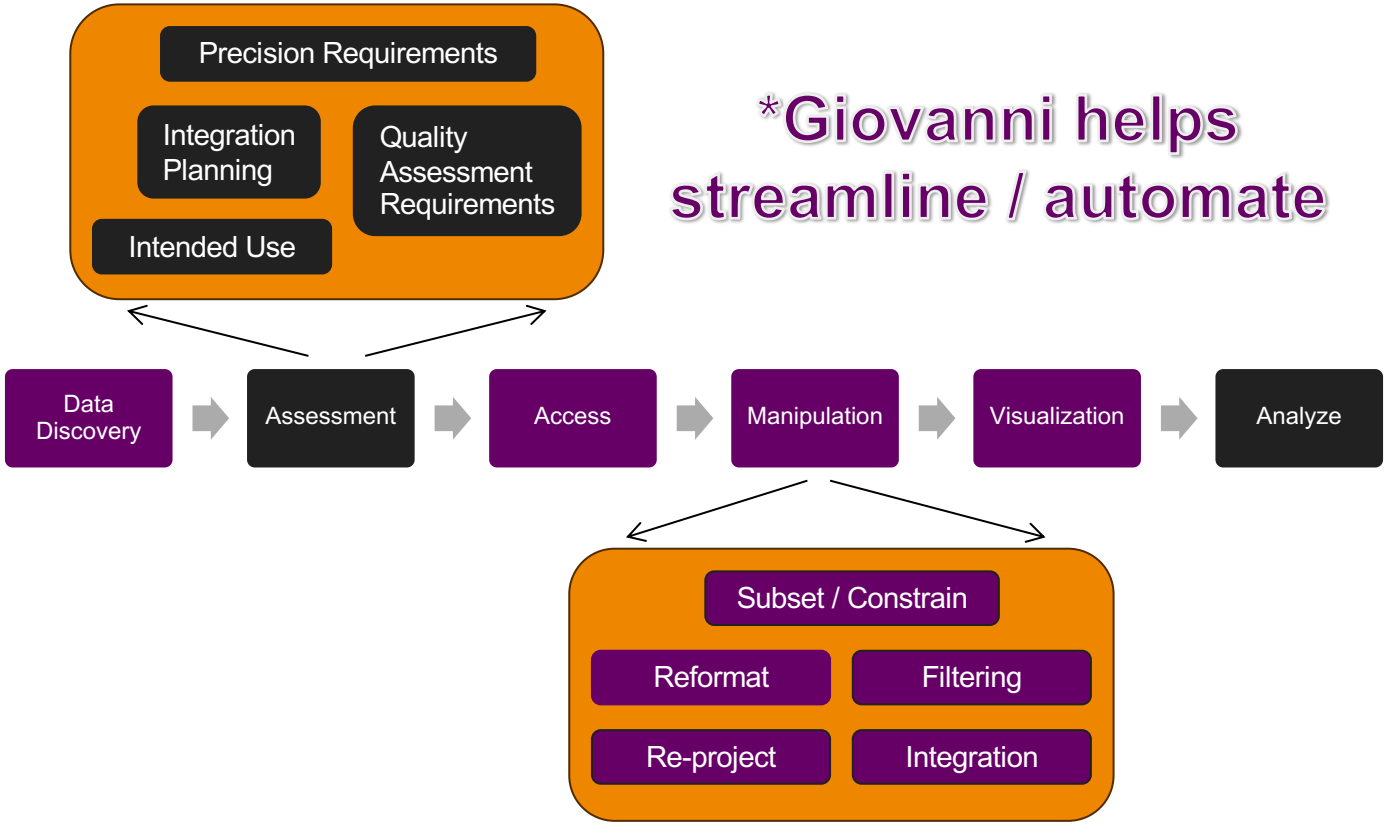


Data Usage Workflow

*Giovanni helps streamline / automate



Data Usage Workflow



Challenges

- Giovanni streamlines data processing, performing required actions on behalf of the user
    - *but* automation amplifies the potential for users to generate and use results they do not fully understand
- The assessment stage is integral for the user to understand fitness-for-use of the result
    - *but* Giovanni does not assist in assessment
- We are challenged to instrument the system to help users understand results

Data Science approach

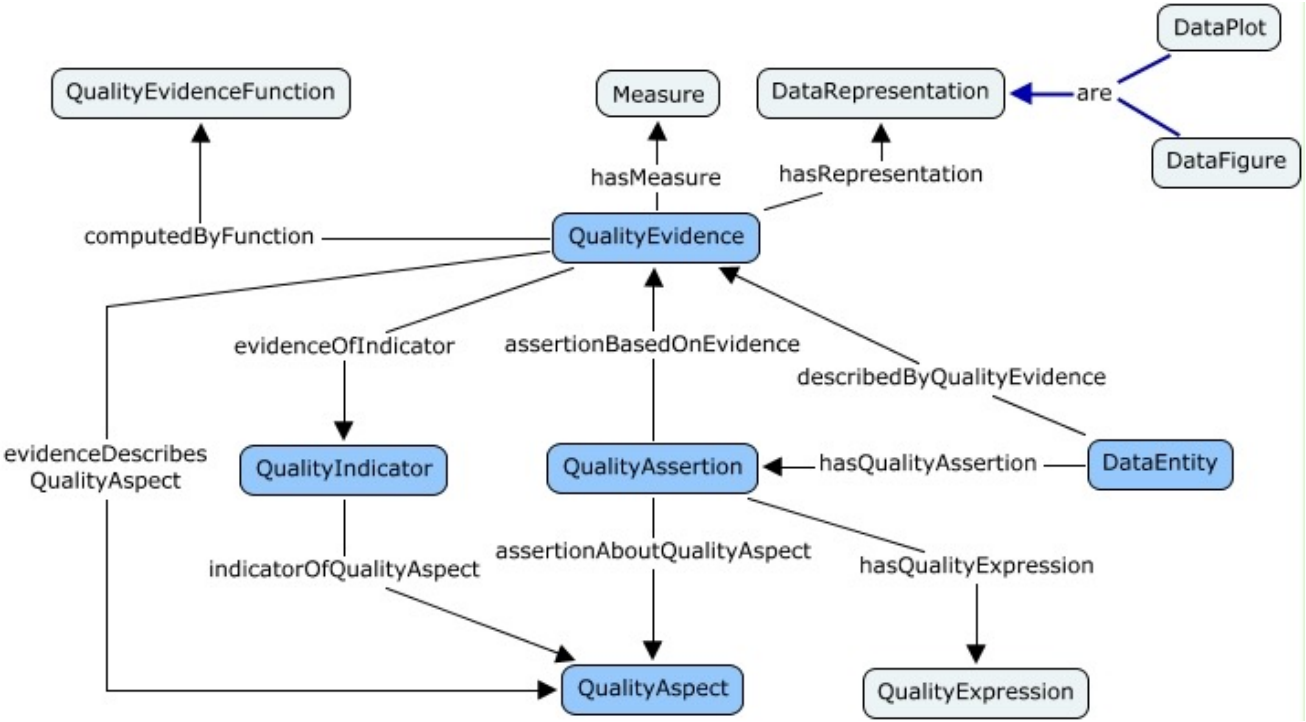
- Systematizing quality aspects
  - Working through literature
  - Identifying aspects of quality and their dependence on measurement and environmental conditions
  - Developing Data Quality encodings
  - Understanding and collecting internal and external provenance
- Developing rulesets allows us to infer pieces of knowledge to extract and assemble
- Presenting data quality knowledge with good visual, statement and references

Advising users on Product quality

We need to be able to advise MDSA users on:

- Which product is better? Or, a better formulated question: Which product has better quality over certain areas?
- Address harmonization of quality across products
- How does sampling bias affect product quality?
  - Spatial: sampling polar area more than equatorial
  - Temporal: sampling one time of a day only
  - Vertical: not sensitive to a certain part of the atmosphere thus emphasizing other parts
  - Pixel Quality : filtering by quality may mask out areas with specific features
  - Clear sky: e.g., measuring humidity only where there are no clouds may lead to dry bias
  - Surface type related issues

Model for Quality Evidence



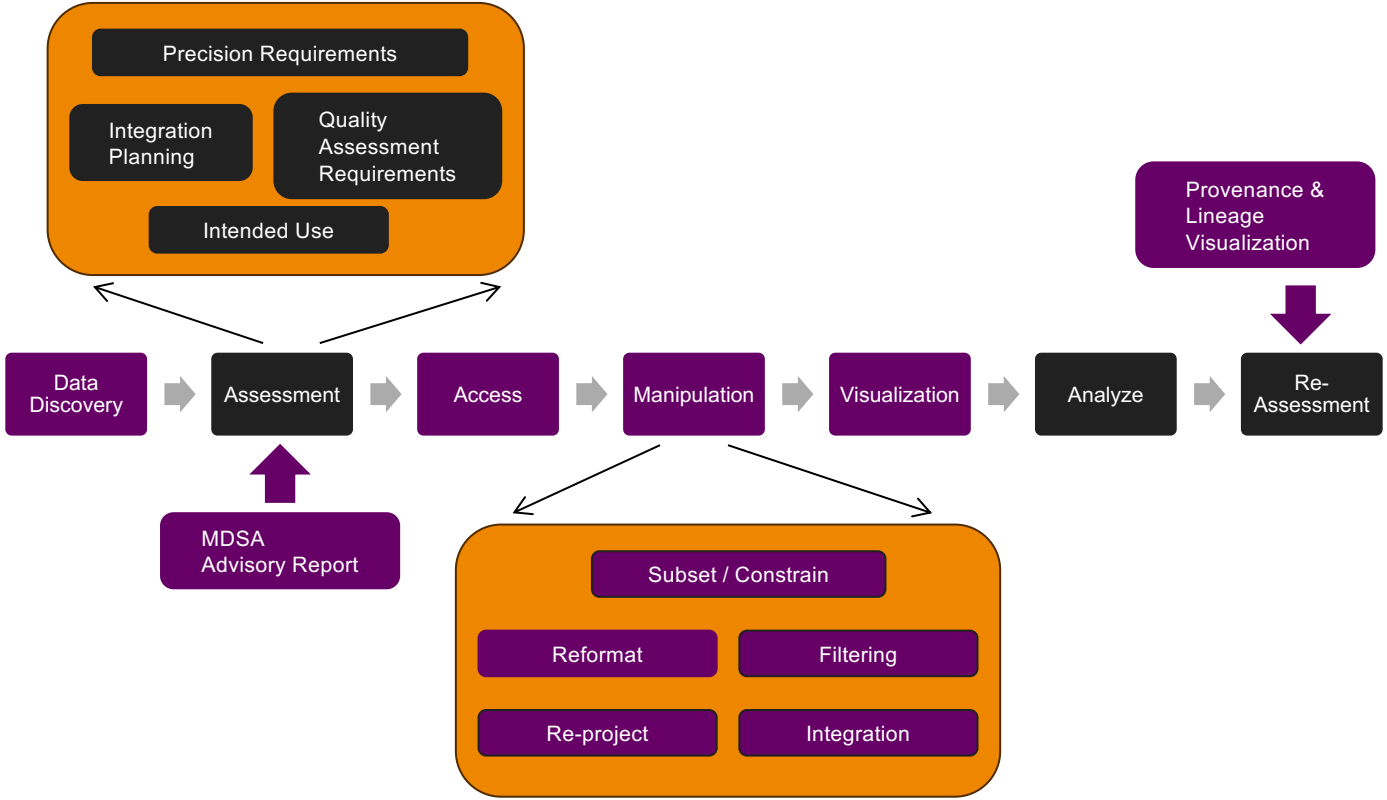
Provenance Distance Computation

Provenance:
Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility.

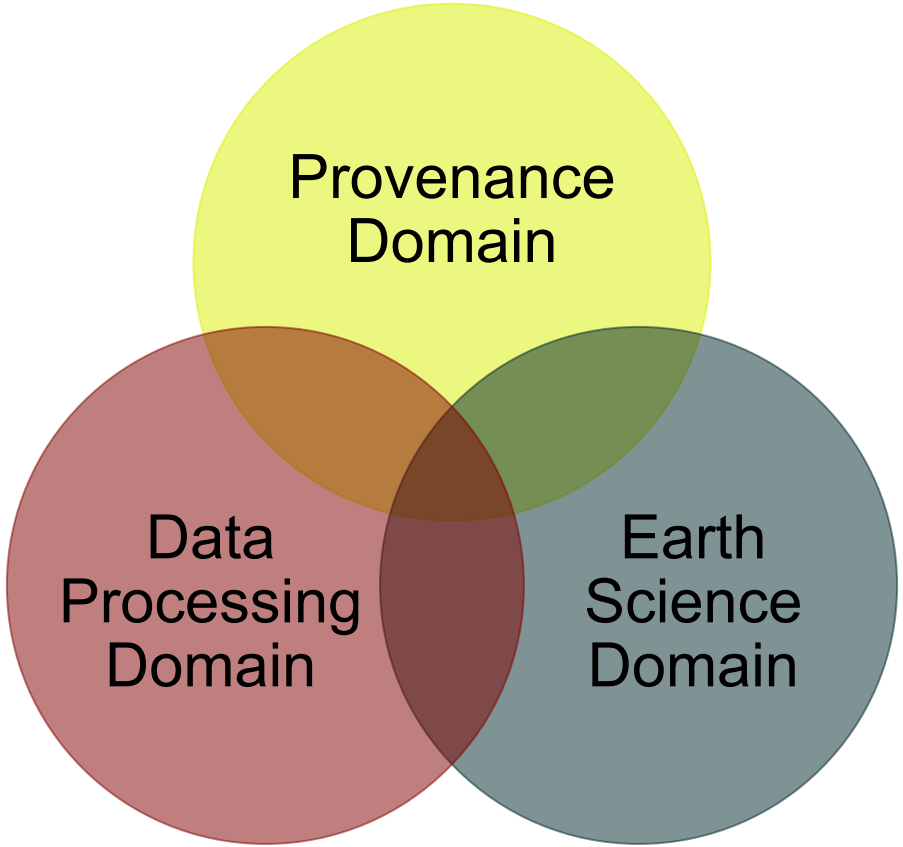
Based on provenance "distance", we tell users how different data products are.
Issues:

- Computing the similarity of two provenance traces is non-trivial
  - Factors in provenance have varied weight on how comparable results of processing are
  - Factors in provenance are interdependent in how they affect final results of processing
- Need to characterize similarity of external (pre-Giovanni) provenance
- Dimensions/factors that affect comparability is quickly overwhelming
- Not all of these dimensions are independent - most of them are correlated with each other.
- Numerical studies comparing datasets can be used, when available, and where applicable to Giovanni analysis

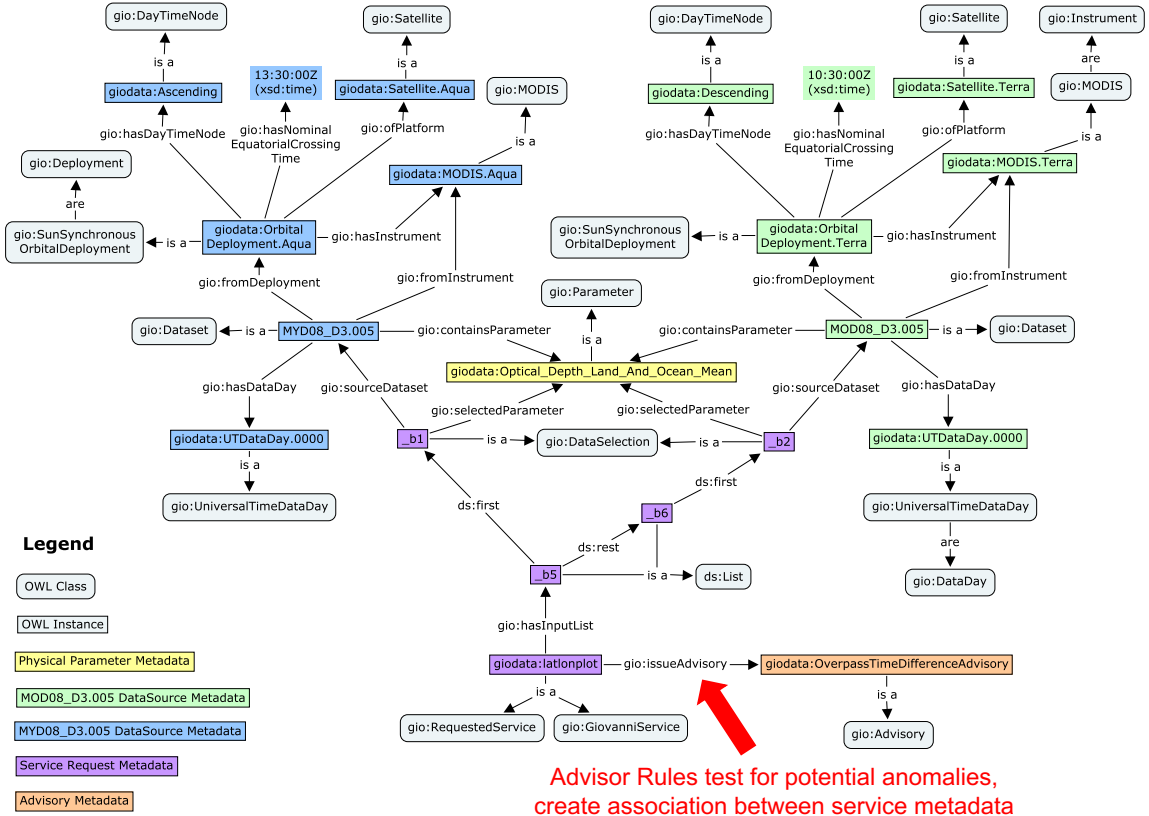
Assisting in Assessment



Multi-Domain Knowledgebase – intersections are for quality



Advisor Knowledge Base



Advisor Rules test for potential anomalies, create association between service metadata and anomaly metadata in Advisor KB

Presenting data quality to users

We split quality (viewed here broadly) into two categories:

- Global or product level quality information, e.g. consistency, completeness, etc., that can be presented in a tabular form.
- Regional/seasonal. This is where we've tried various approaches:
    - maps with outlines regions, one map per sensor/parameter/season
    - scatter plots with error estimates, one per a combination of Aeronet station, parameter, and season; with different colors representing different wavelengths, etc.

Advisor Presentation Requirements

- Present metadata that can affect fitness for use of result
- In comparison or integration data sources
  - Make obvious which properties are comparable
  - Highlight differences (that affect comparability) where present
- Present descriptive text (and if possible visuals) for any data usage caveats highlighted by expert ruleset
- Presentation must be understandable by Earth Scientists!!

Summary

- Quality is very hard to characterize, different groups will focus on different and inconsistent measures of quality
    - Modern ontology representations to the rescue!
- Products with known Quality (whether good or bad quality) are more valuable than products with unknown Quality.
    - Known quality helps you correctly assess fitness-for-use
- Harmonization of data quality is even more difficult that characterizing quality of a single data product

Acronyms

ACCESS	Advancing Collaborative Connections for Earth System Science
ACE	Aerosol-Cloud-Ecosystems
AGU	American Geophysical Union
AIST	Advanced Information Systems Technology
AOD	Aerosol Optical Depth
AVHRR	Advanced Very High Resolution Radiometer
GACM	Global Atmospheric Composition Mission
GeoCAPE	Geostationary Coastal and Air Pollution Events
GEWEX	Global Energy and Water Cycle Experiment
GOES	Geostationary Operational Environmental Satellite
GOME-2	Global Ozone Monitoring Experiment-2
JPSS	Joint Polar Satellite System
LST	Local Solar Time
MDSA	Multi-sensor Data Synergy Advisor
MISR	Multiangle Imaging SpectroRadiometer
MODIS	Moderate Resolution Imaging Spectraradiometer
NPP	National Polar-Orbiting Operational Environmental Satellite System Preparatory Project

Acronyms (cont.)

OMI	Ozone Monitoring Instrument
OWL	Web Ontology Language
PML	Proof Markup Language
QA4EO	QA for Earth Observations
REST	Representational State Transfer
TRL	Technology Readiness Level
UTC	Coordinated Universal Time
WADL	Web Application Description Language
XML	eXtensible Markup Language
XSL	eXtensible Stylesheet Language
XSLT	XSL Transformation

Thanks!

Meet with your teammates!