Input for NASA/NRC Information Systems and Processing Technologies: data management and processing

Peter Fox

Tetherless World Constellation Chair, Prof. of Earth and Environmental Science and Computer Science, Rensselaer Polytechnic Institute

Table 2, p. TA11-20 and …

- Most of this looks right on track, except perhaps a format error in the Search row?
- The declared separation of computation/ tools/ data is contrary to the direction being set/taken elsewhere
  - I.e. – data close to computation, tools close to user
- Sec. 2.2.4.3 misses much of the VO/STS experience (ask me if you want details)
- The remainder of the draft covers a lot of good material – but you asked about GAPS

Gaps …

- Data integration or integrate-ability…
- Too far off adoption of semantic technologies
  - When they are already in widespread production in NASA (GSFC, Ames, JPL, Langley)
- TA11-23 - Representing and Reasoning about Complex States and Resources; Spatial reasoning: and no mention of semantics?
- Fitness … Quality, uncertainty, bias
- NASA Earth Science Technology Office infusion and roadmaps – no mention of them – semantics, sensors, web services, provenance (obs/ and models)

Where are we in respect to data challenges?

*"The user cannot **find** the data;*

*If he can find it, cannot **access** it;*

If he can access it, ;

*he doesn't know **how good** they are;*

*if he finds them good, he can not **merge** them with other data"*

The Users View of IT, NAS 1989

Fitness for purpose: Quality +

- **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**

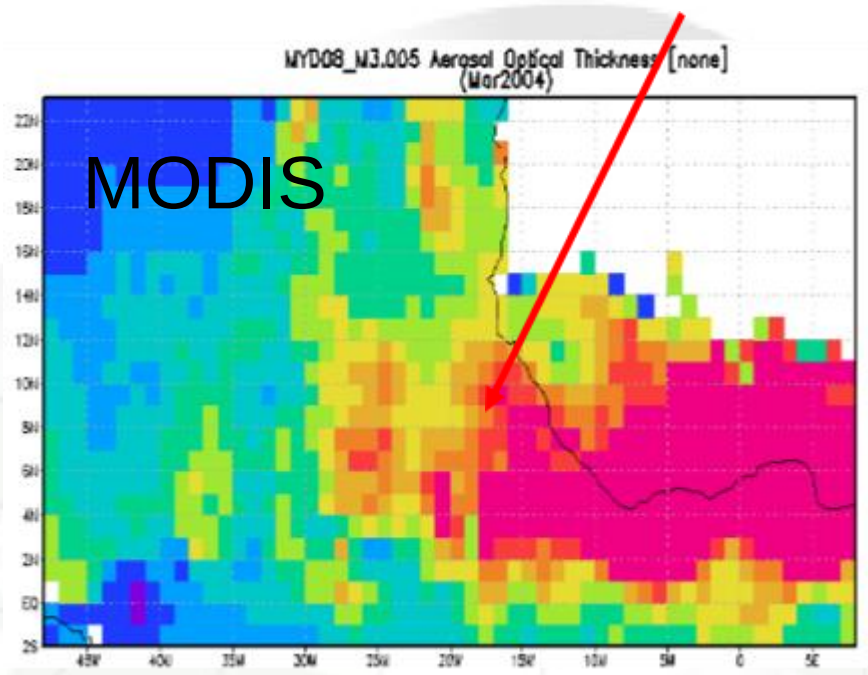
However NO mention of ..

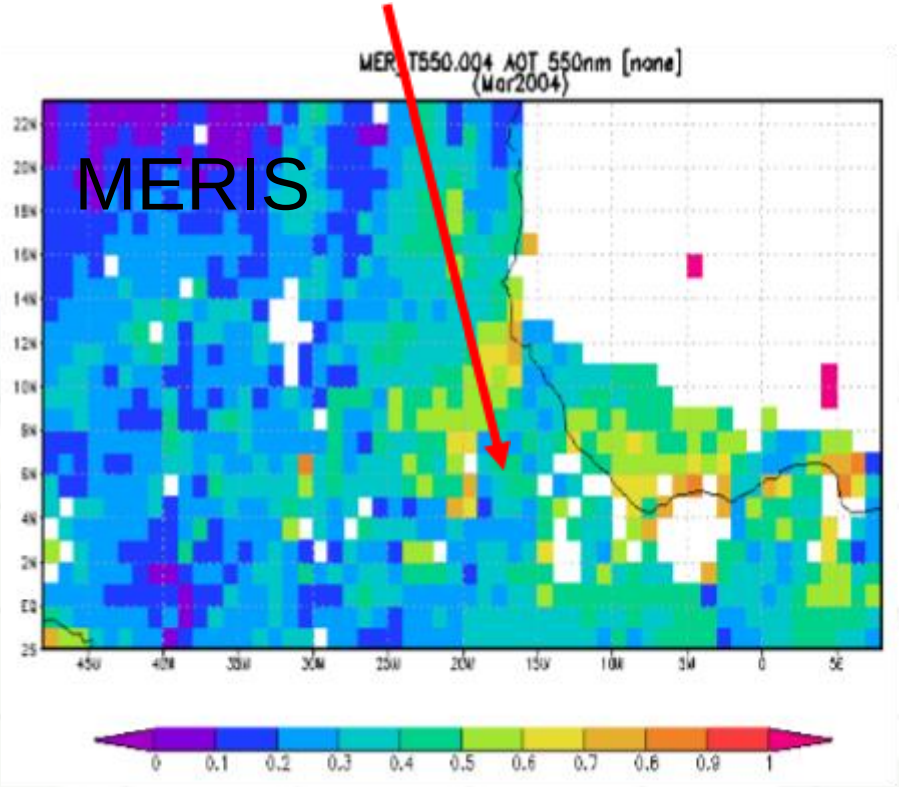
- Quality
  - Very much in the eyes of the beholder – worst case scenario… or a good challenge OR declared by the originator – big gap
- Uncertainty
  - has aspects of accuracy (how accurately the real world situation is assessed, it also includes bias) and precision (down to how many digits)
- Bias 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)

E.g. AOT MODIS vs. MERIS

Same parameter

Same space & time



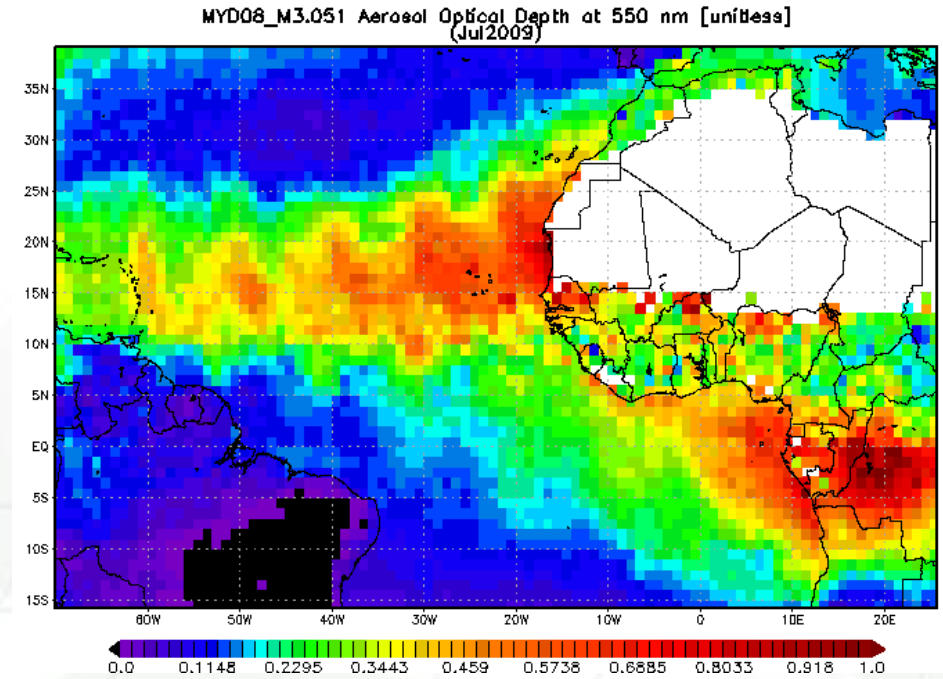


Different results – why?

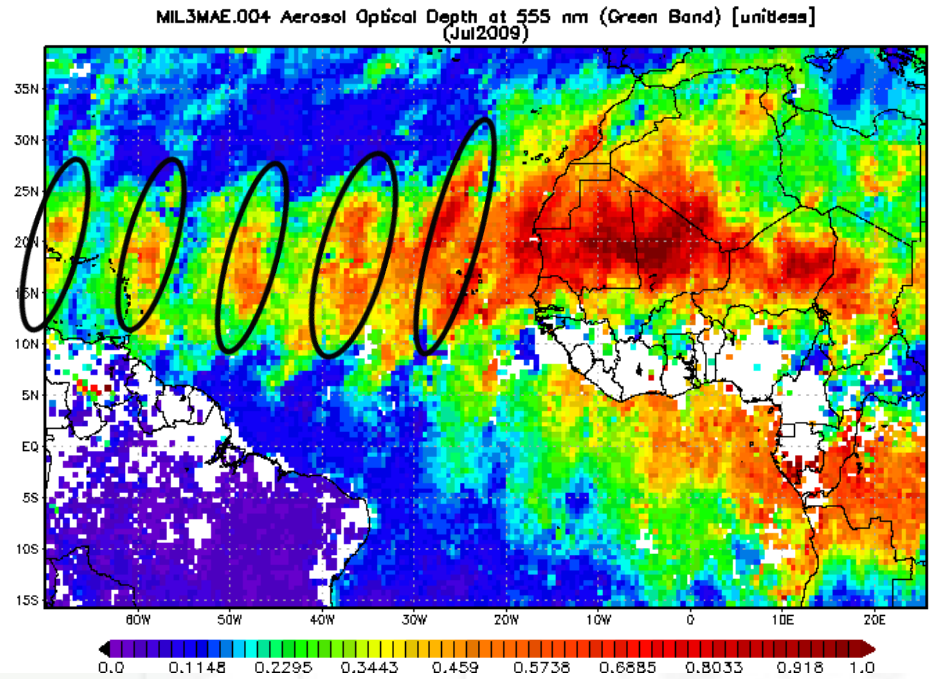
A threshold used in MERIS processing effectively excludes high aerosol values. *Note: MERIS was designed primarily as an ocean-color instrument, so aerosols are "obstacles" not signal.*

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

MODIS Aqua AOD July 2009

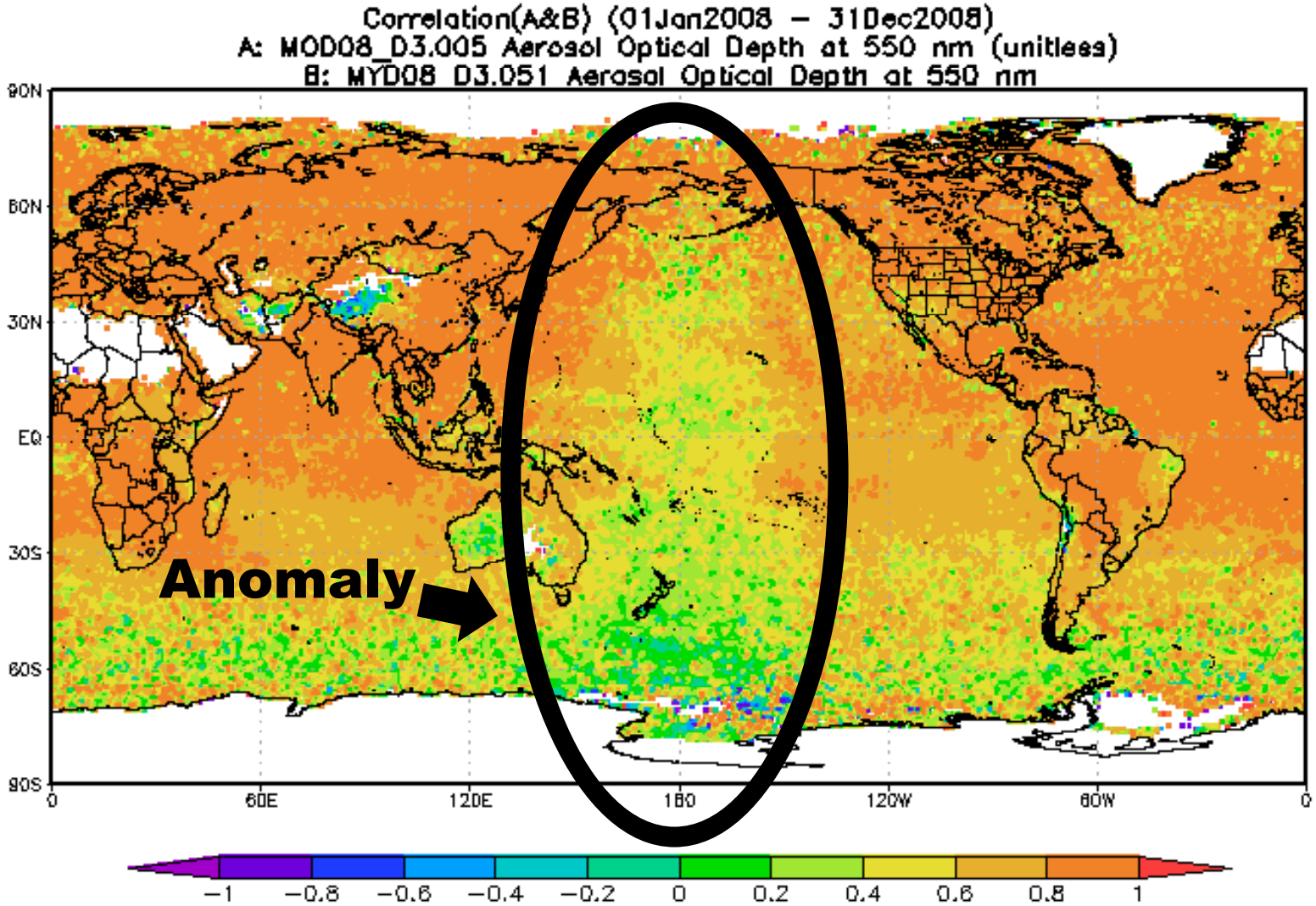


MISR Terra AOD July 2009



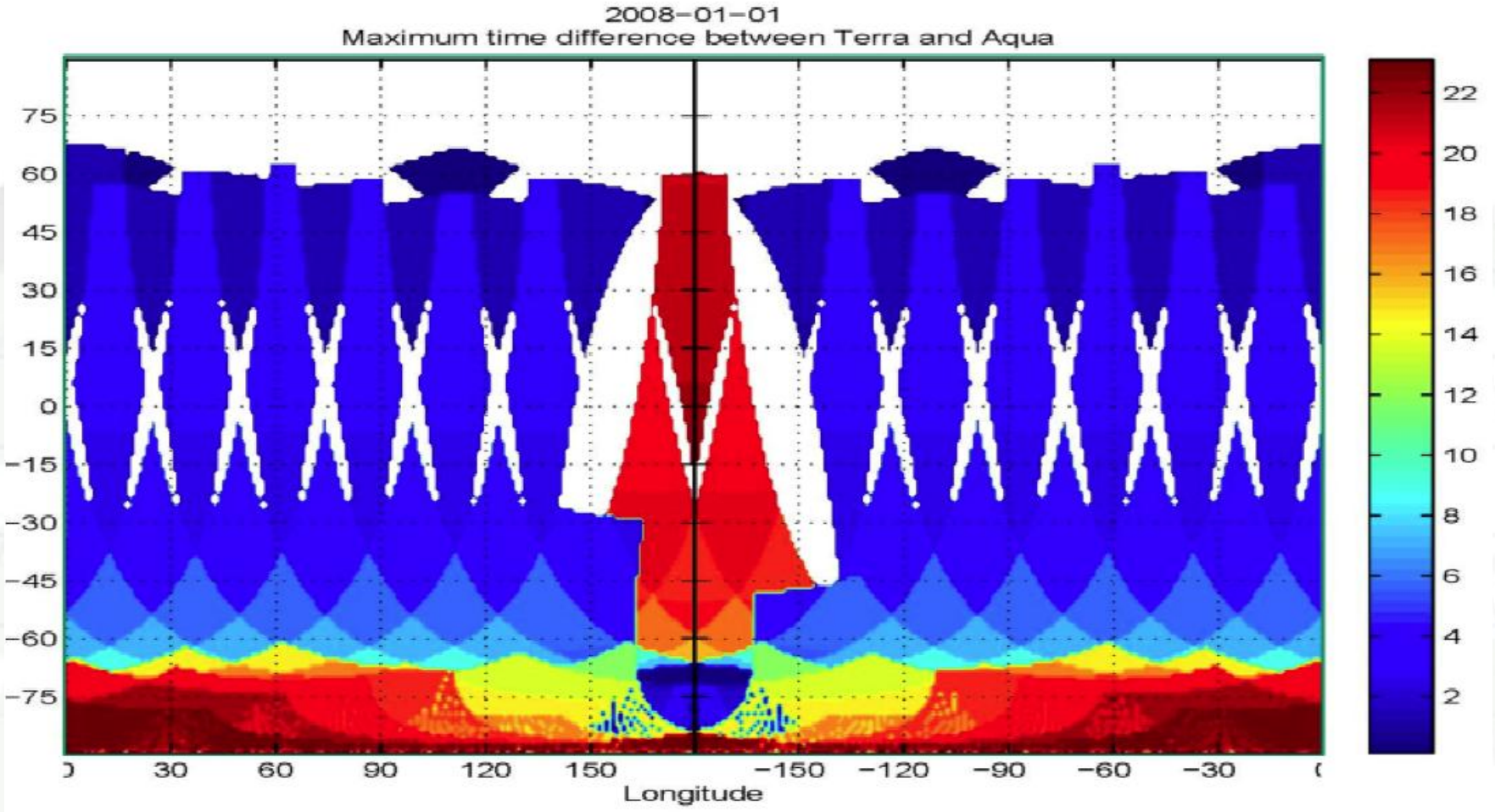
- *Completeness*: MODIS dark target algorithm does not work for deserts
- *Representativeness*: monthly aggregation is not enough for MISR and even MODIS
- *Spatial sampling* patterns are different for MODIS Aqua and MISR Terra: "pulsating" areas over ocean are oriented differently due to different orbital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

Anomaly Example: South Pacific Anomaly

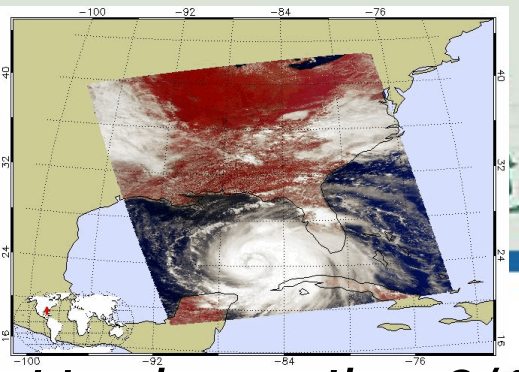


MODIS Level 3 dataday definition leads to artifact in correlation

…is caused by an Overpass Time Difference

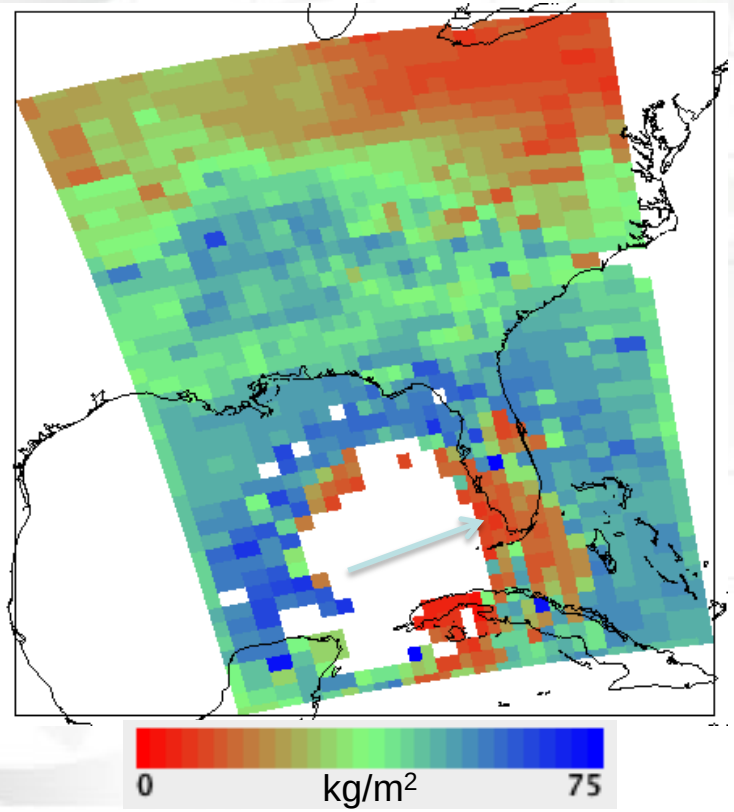


The effect of bad quality data is often *not* negligible

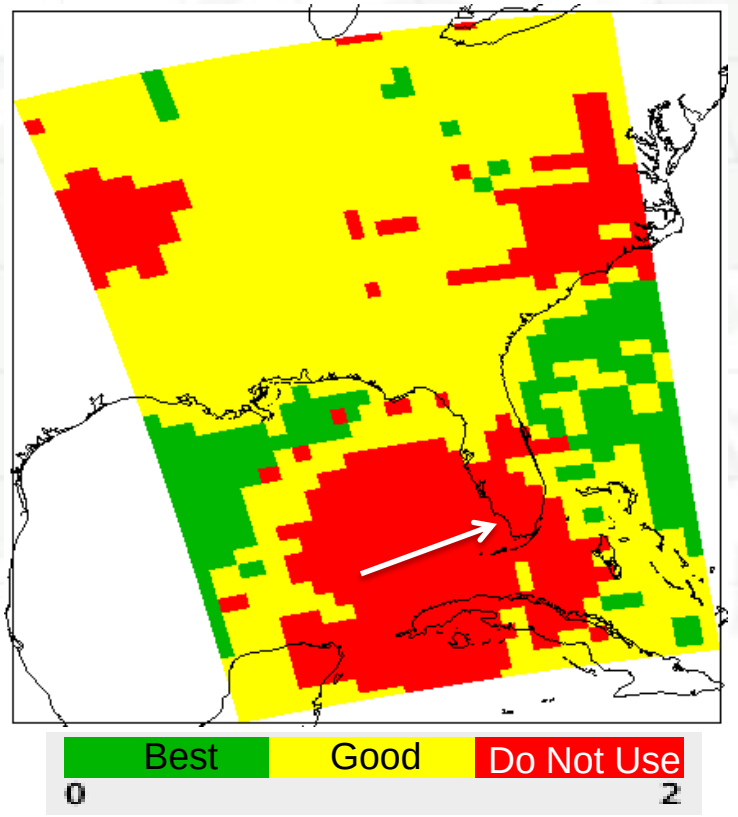


Hurricane Ike, 9/10/2008

Total Column Precipitable Water



Quality



E.g. Giovanni Allows Scientists to Concentrate on the *Science*

The Old Way:

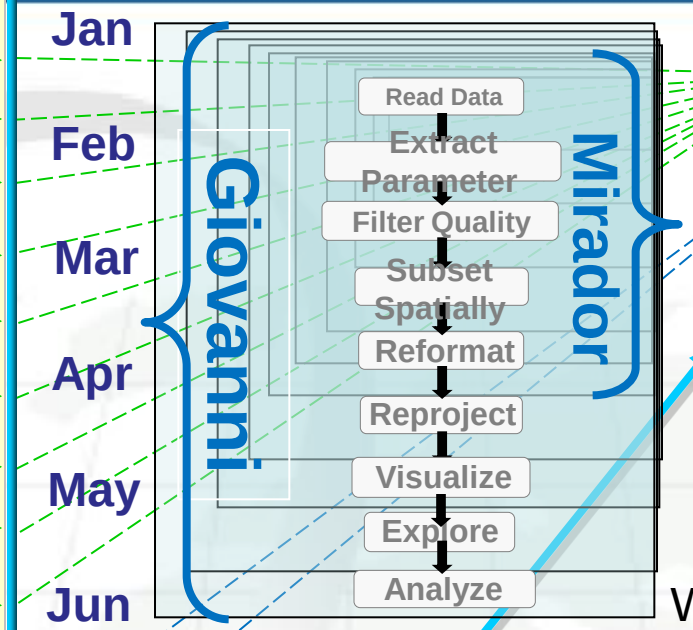
Pre-Science

- 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:



The Giovanni Way:

- Minutes
- Days for exploration

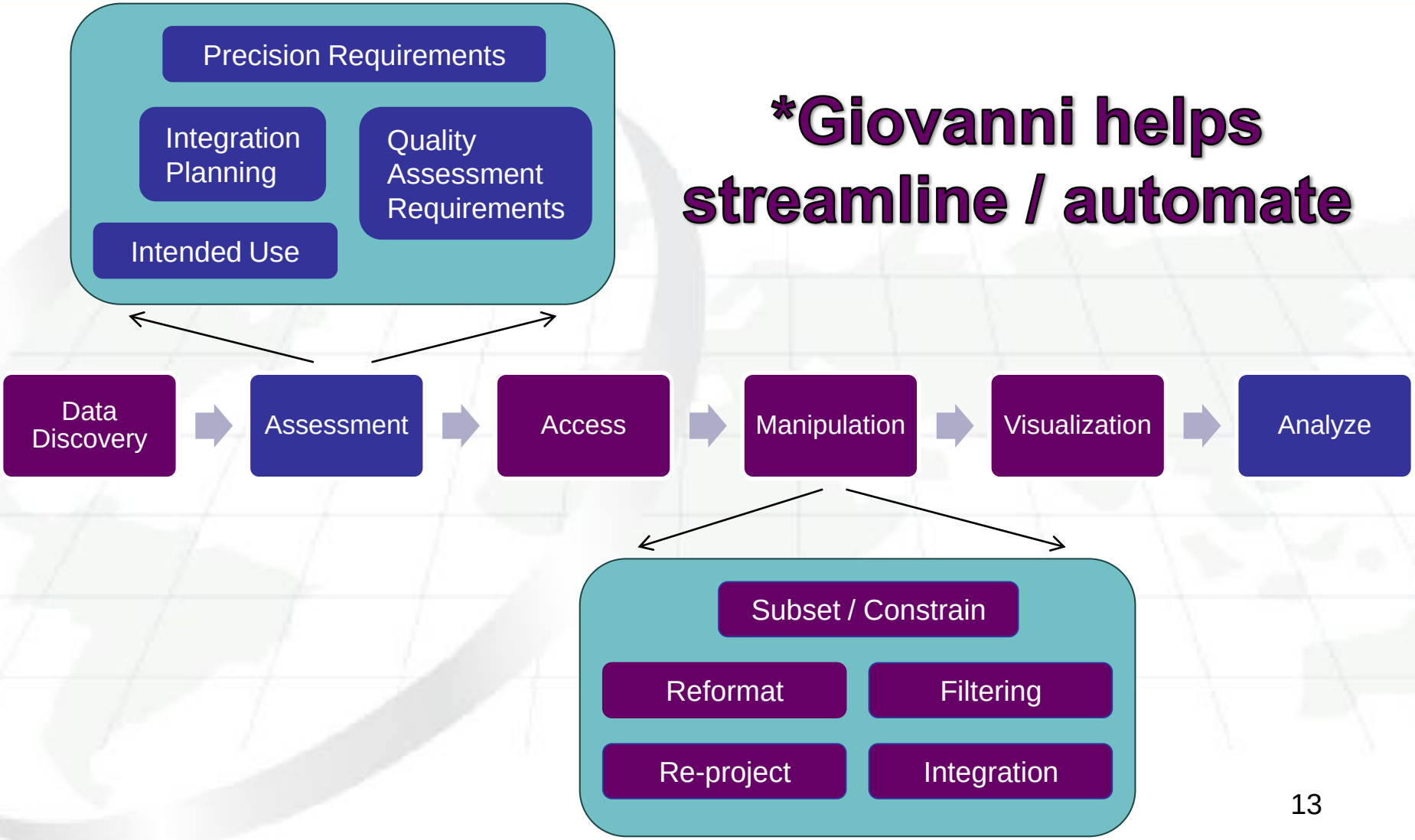
DO SCIENCE

- Use the best data for the final analysis
- Derive conclusions
- Write the paper
- Submit the paper

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*.

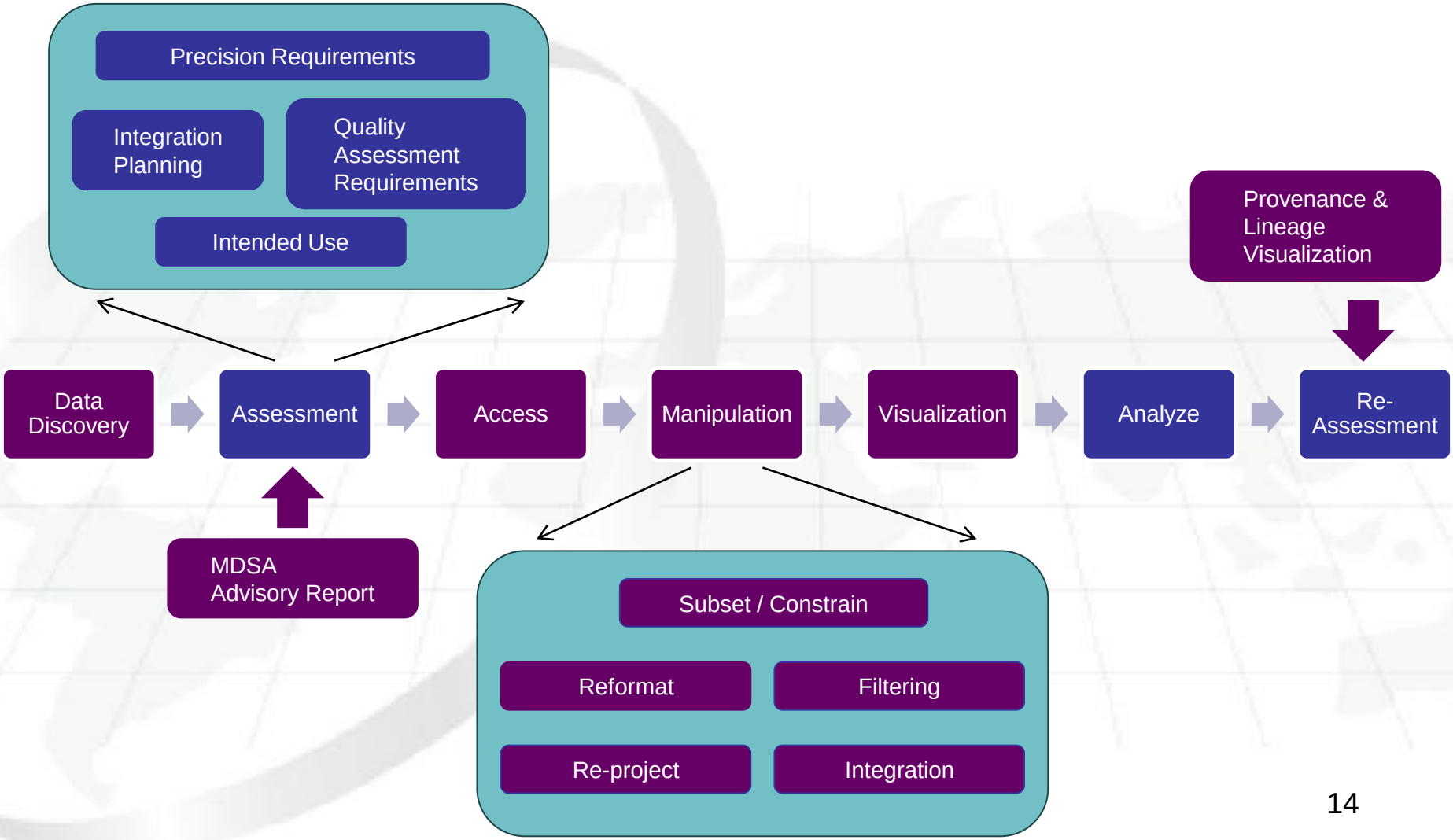
Scientists have ***more time to do science!***

Data Usage Workflow



***Giovanni helps streamline / automate**

Assisting in Assessment

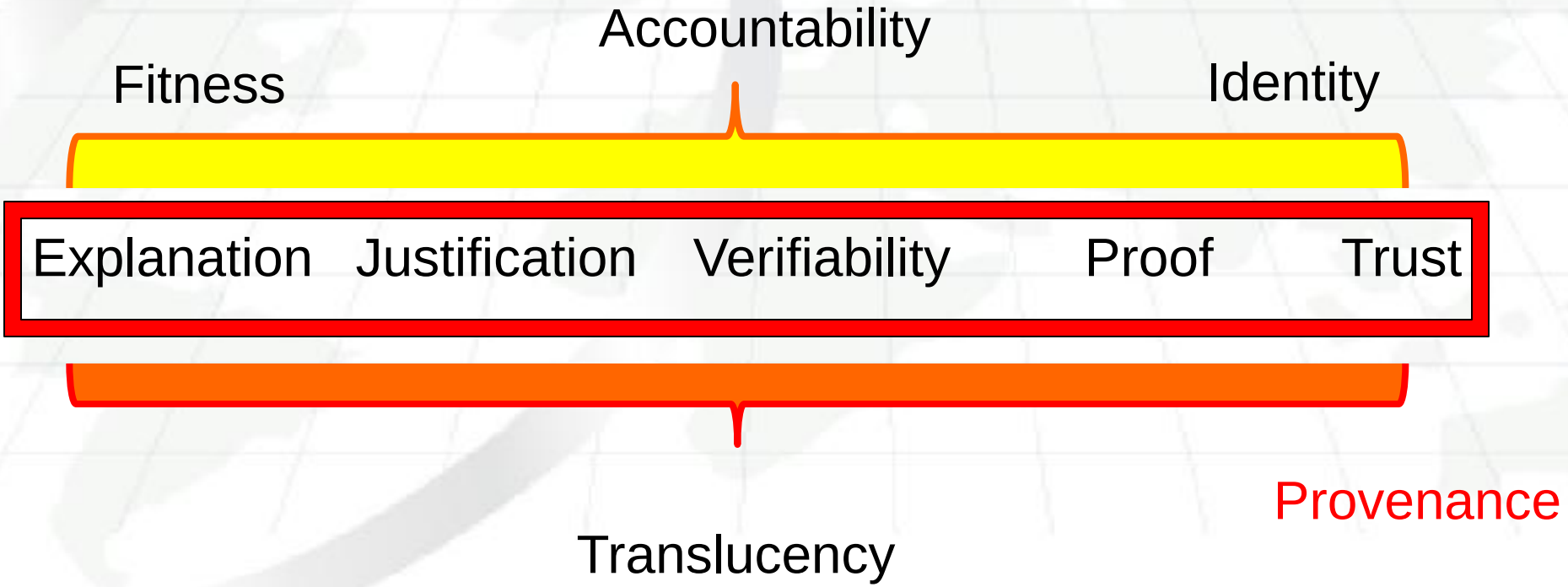


Semantic Web Roadmap

		Current	Near Term (0-2 yrs)	Mid Term (2-5 yrs)	Long Term (5+ yrs)
Results	Outcome	♦ Improved Information Sharing	♦ Increased Collaboration & Interdisciplinary Science	♦ Acceleration of Knowledge Production	♦ Revolutionizing how science is done
	Output	♦ Geospatial semantic services established	♦ Geospatial semantic services proliferate	♦ Scientific semantic assisted services	♦ Autonomous inference of science results
Capability	Assisted Discovery & Mediation	♦ Some common vocabulary based product search and access	♦ Semantic geospatial search & inference, access	♦ Semantic agent-based searches	♦ Semantic agent-based integration
	Interoperable Information Infrastructure	♦ Local processing + data exchange	♦ Basic data tailoring services (data as service), verification/ validation	♦Interoperable geospatial services (analysis as service), results explanation service	♦ Metadata-driven data fusion (semantic service chaining), trust
Technology	Vocabulary	♦ SWEET core 1.0 based on GCMD/CF	♦ SWEET core 2.0 based on best practices decided from community	♦ SWEET 3.0 with semantic callable interfaces via standard programming languages	♦ Reasoners able to utilize SWEET 4.0
	Languages/ Reasoning	♦ RDF, OWL, OWL-S	♦ Geospatial reasoning, OWL-Time	♦ Numerical reasoning	♦ Scientific reasoning

Info. processing ecosystem?

- These are what enable scientists to explore/ confirm/ deny their ‘hunches’



Thanks

- pfox@cs.rpi.edu