Improving Interoperability and Science Productivity: *Proposal for A Data Analysis Environment*

K.-S. Kuo

GEST/Caelum Research Corp.

NASA Goddard Space Flight Center

Background

- I am a scientist user with extensive experience in writing code, but not a technologist.

MOTIVATIONS (1/2)

- Experience with the creation of CloudSat-TRMM data product.
  - Combine near-coincident observations of two spaceborne radars, *i.e.* CloudSat's CPR at 94 GHz and TRMM's PR at 14 GHz
  - CPR is sensitive to smaller cloud hydrometeor particles whereas PR is sensitive to larger precipitation particles.
  - Together they provide a more comprehensive view of the cloud/precipitation system.
- Conclusion: ***Very difficult!***
  - Finding the orbit crossings
  - Finding the coincident portions in both data sets
  - Combining them into a new HDF (4) data file

MOTIVATIONS (2/2)

- Combining data products from the same instrument but different disciplines (*e.g.* MODIS aerosol, MODIS cloud, and/or MODIS land) is much easier but still difficult.
- Can researchers take advantage of what others had already done?
- There are Integrated Development Environments (IDEs) that integrate all relevant tools and handle all aspects of software development, including editing, compiling, debugging, testing, versioning, *etc.*, *i.e.* the whole life cycle.
- Can we create some Integrated Data Environments for Analysis (IDEA) that integrate all relevant tools and handle all aspects of data analysis? (Obviously, this is much more ambitious than IDEs.)

Wish List (1/2)

- Platform independent – no favoritism to MS
- Language indifferent – new language skill not required
- An underlying data model (OPeNDAP?) appropriate for remote sensing data
- Conducive to and compatible with web services
- Client-Server – remote access, unlike Eclipse
- Tool and service repository/catalogs – reuse with confidence

Wish List (2/2)

- Easy integration of new tools – facilitating exploration and expansion
- Guidelines for user interface – consistent and familiar look-and-feel
- Workflow construction
- Essential capabilities first – maximizing impact early on
- Automatic logging – provenance
- Open-source – perpetuating development

FUNCTIONALITIES

Entire data life cycle:

- Data discovery
- Data access
- Data management
- Integration of customized data analysis tools
- Data analysis
- Data generation
- Data publication?

Guiding Principle

- Adapting and integrating existing tools whenever possible
  - Document adaptation/integration process to serve as tutorial/example for others to follow!

Obstacles

- Idiosyncrasies of the instruments and data products
  - Units – Hz, micron, wave number in inverse cm, …
  - Nomenclatures – optical thickness or depth?
  - Conventions – larger number for better or worse quality?
  - Disparity in resolutions – How to compare or combine them?
- Metadata and conversion/translation?
- User preferences?

Test Bed

- Can we demonstrate what we can accomplish now with a prototype using a subset of data products?

THANK YOU!

Barriers to Seamless Interoperation

- Resource/Financial
- Scientific and Technological
  - Radiation is multi-directional (multiple vectors) at any given position, sort of like wind (unidirectional, one vector) but more complicated.