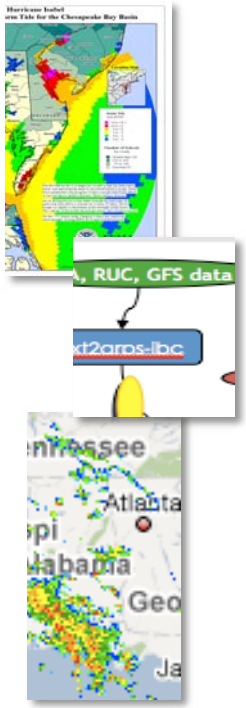


Metadata and Provenance: Dimensions of Use in Science

Beth Plale

Director, Data To Insight Center

Professor of Computer Science

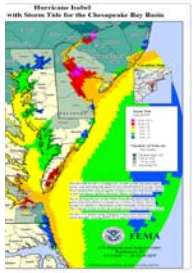
Managing Director, Pervasive Technology Institute

Co-PI Sustainable Environments – Actionable Data (SEAD) NSF DataNet

Indiana University, USA

DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute

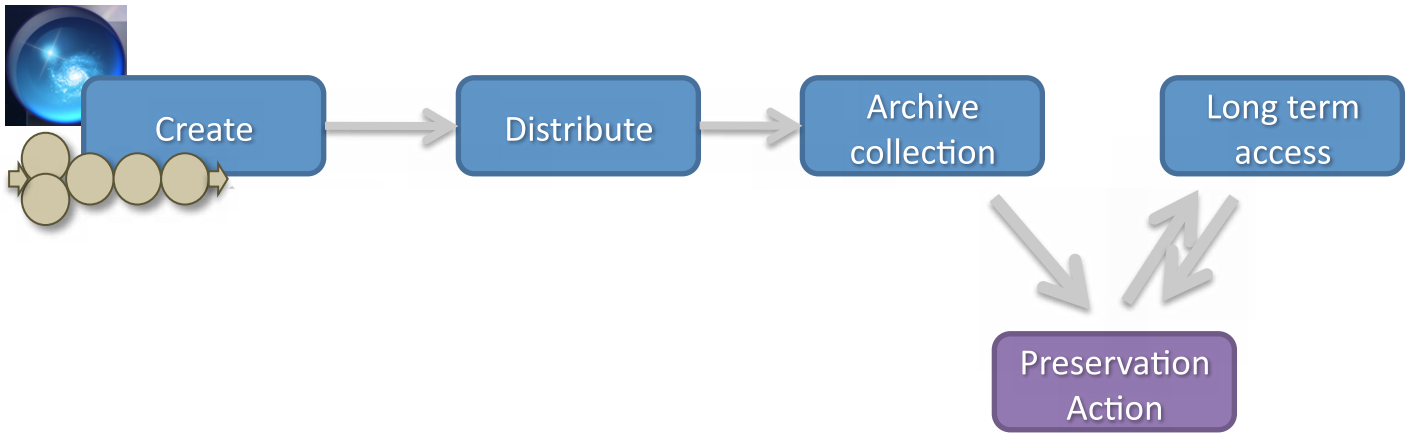
PRAGMA April 2012




Challenges of Science Data Data Deluge

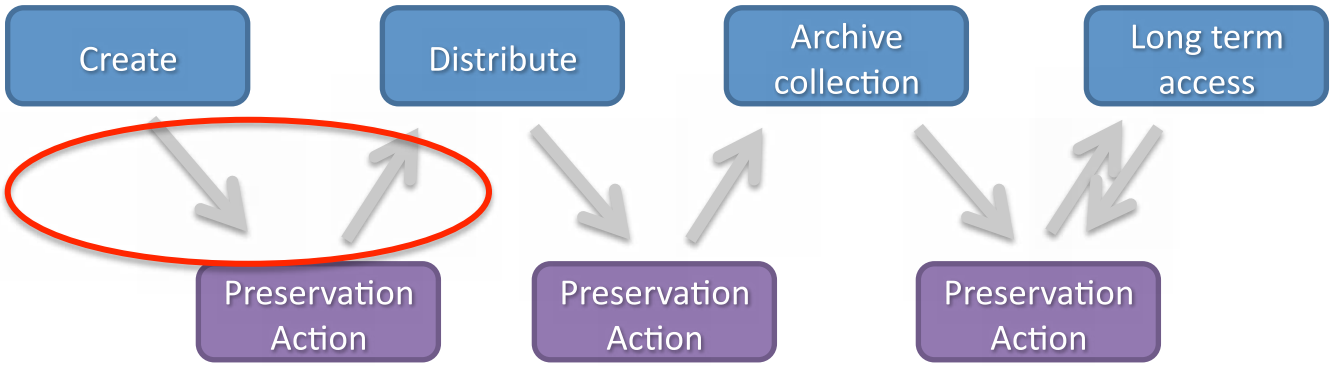
- Metadata must be preserved when scientific data is generated because metadata is ephemeral – *Jim Gray*
- If annotation is left to the scientist, it is not done (U.K. e-Science Core)
- The further the distance between data producer and re-use, the more detailed the metadata that's required.

Typical Data Lifecycle



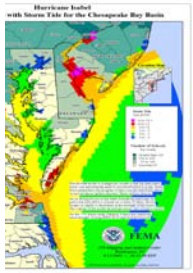
- Problem: metadata capture too late. Use of scientific data 5-50 years from requires that metadata capture occur early in life of data otherwise curation costs are too high. Scientists tools have to be enabled to help with preservation in earliest stages of data's life.

Goal or Objective Data Lifecycle



- Tools for early curation: capture information for both Discovery AND Use
- Ontologies and metadata should be inextricably linked

Diagram from Berman et al. "Sustainable Economics for a Digital Planet"




Discovery

- The legacy solution to discovering data is to embed lots of metadata into file names
  - http://lead.unidata.ucar.edu:8080/thredds/dodsC/LEAD/radar2/KVTX/20090914/Level2_KVTX_20090914_1321.ar2v
  - http://lead.unidata.ucar.edu:8080/thredds/fileServer/LEAD/model/NCEP/NAM/CONUS_80km/NAM_CONUS_80km_20090914_1200.grib1
- Good for those initiated into “inner-circle”
- Relying on long file names isn’t enough

Discovery

Portal

Name: wrfout_d01_2009-03-05_12:00:00
GUID:urn:uuid:b419247e-876d-4842-b463-e79fc50aea3b
Owner:/O=LEAD Project/OU=portal.leadproject.org/
OU=cs.indiana.edu/CN=plale/EMAIL=plale@cs.indiana.edu
Create time: …

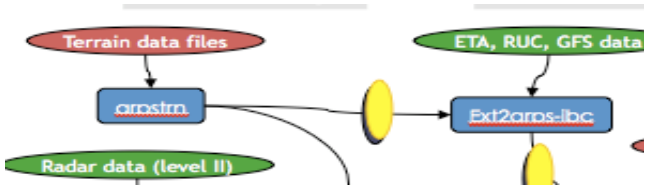
File system (e.g., Data Capacitor)

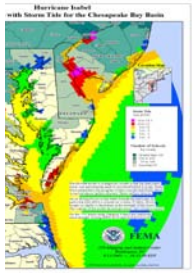
b419247e-876d-4842-b463-e79fc50aea3b

Query: Give me all related data about Mount Kinabalu unltramafic research done 2005-2010

Metadata in database

Objects stored to file system, OPeNDAP, iRods

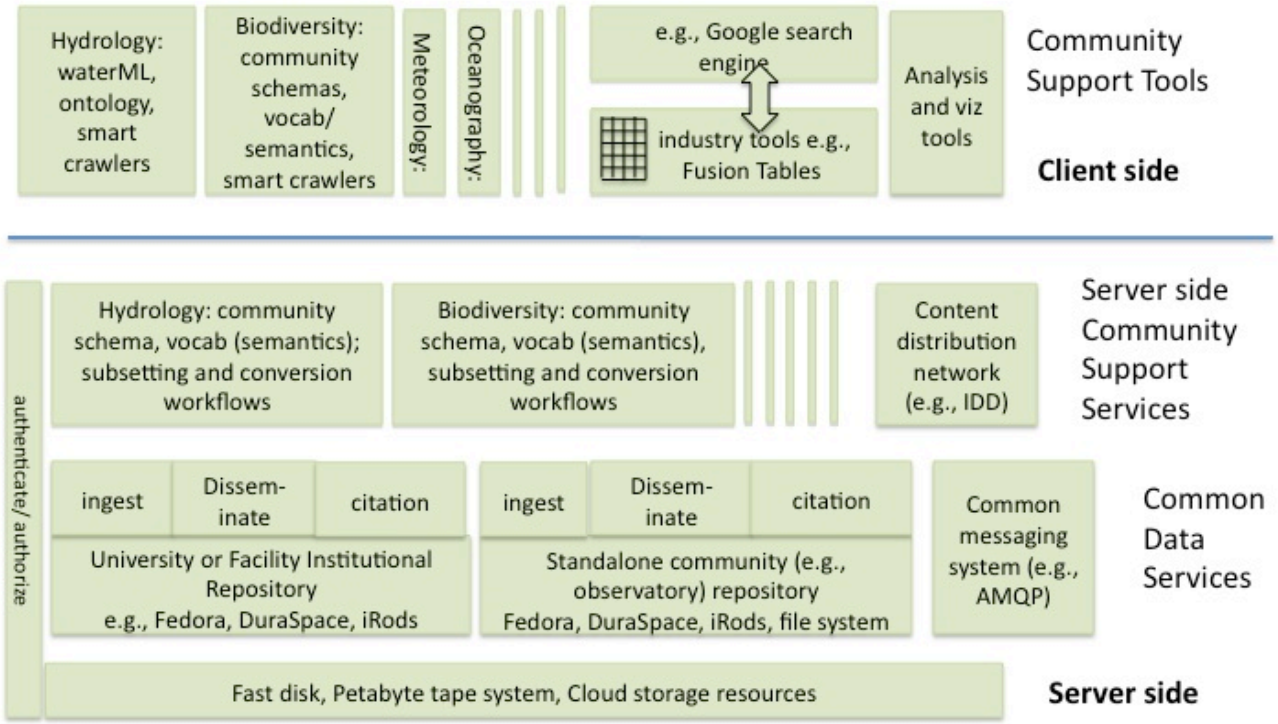




Actionable (Use)

- What does value X mean (in my domain's vocabulary)?
- What coordinate system was used in this model?
- Data set is big, I need only subset over Mount Kinabalu April 2012
- What use restrictions are on this data?
- Who generated this data, and under what conditions?

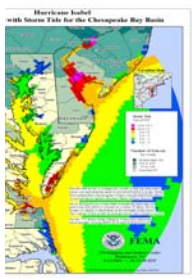
Metadata, Linked Data and Ontologies



Takeaway point: ontologies and metadata schemas complement each other to provide coverage for both discovery AND use metadata

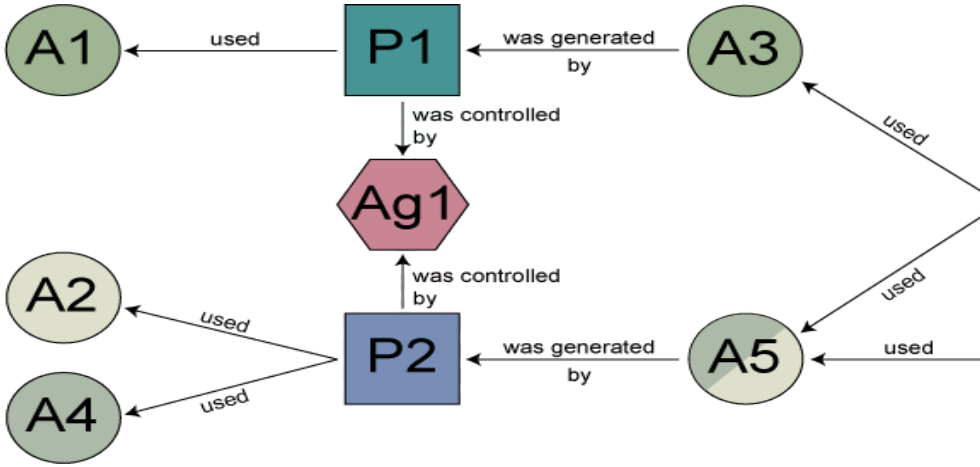
PRAGMA April 2012

Plale et al. Atmospheric Sciences and Informatics EarthCube Driver Whitepaper: Technical Infrastructure, Fall 2011




Data Provenance

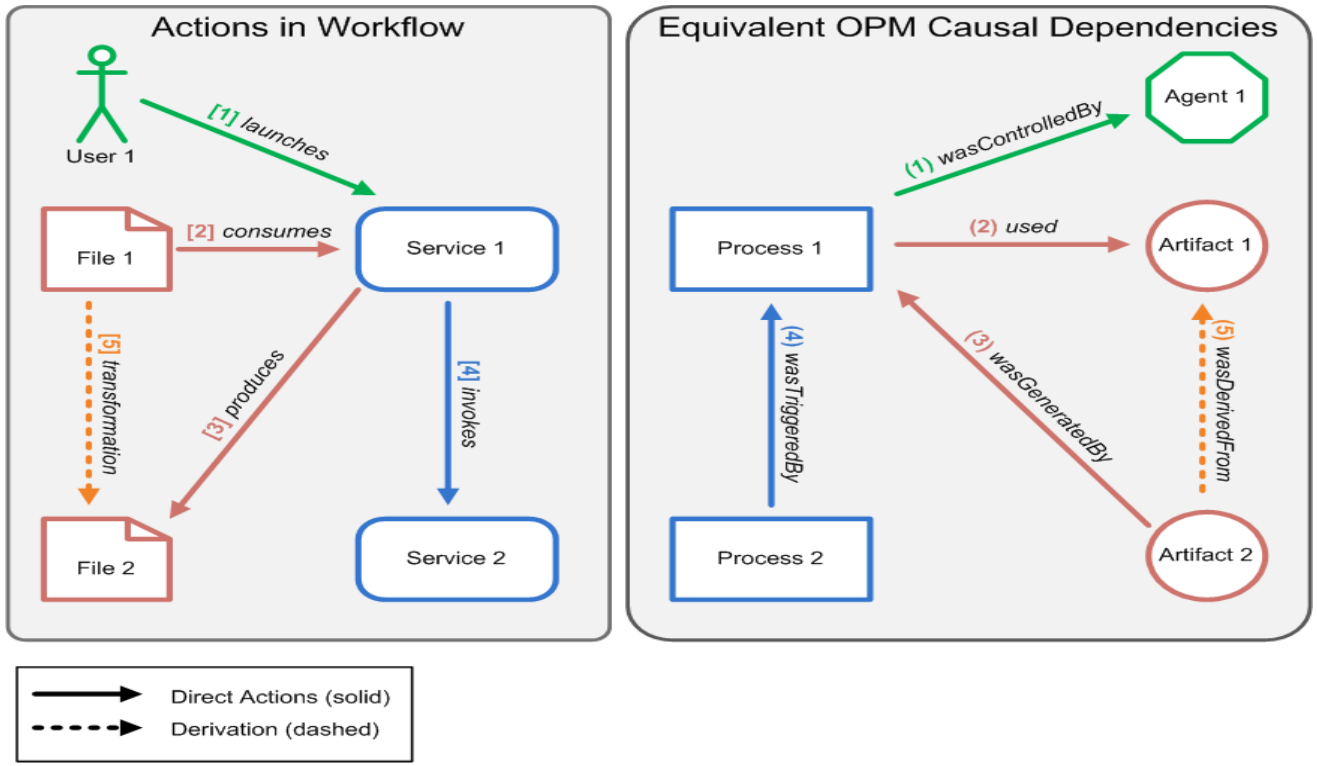
Provenance is lineage of data object or collection.
Explains what contributed to object's creation



Data Provenance: analogy to provenance of works of art

- Trace of history of work of art from moment it was made until it comes into a collection.
- Impartial and authoritative information on authenticity, ownership, theft, and other artistic, legal, and ethical issues concerning art objects.

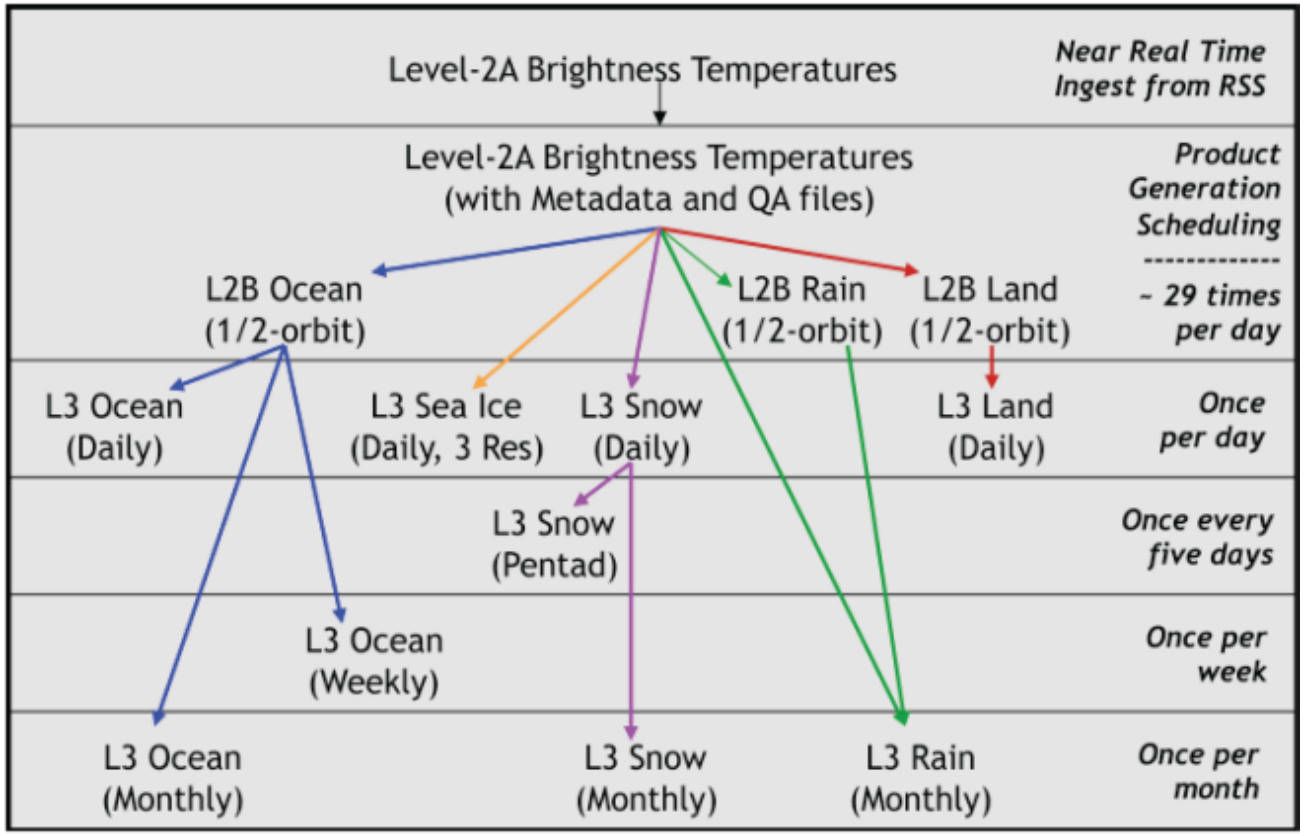
Types of Provenance Information



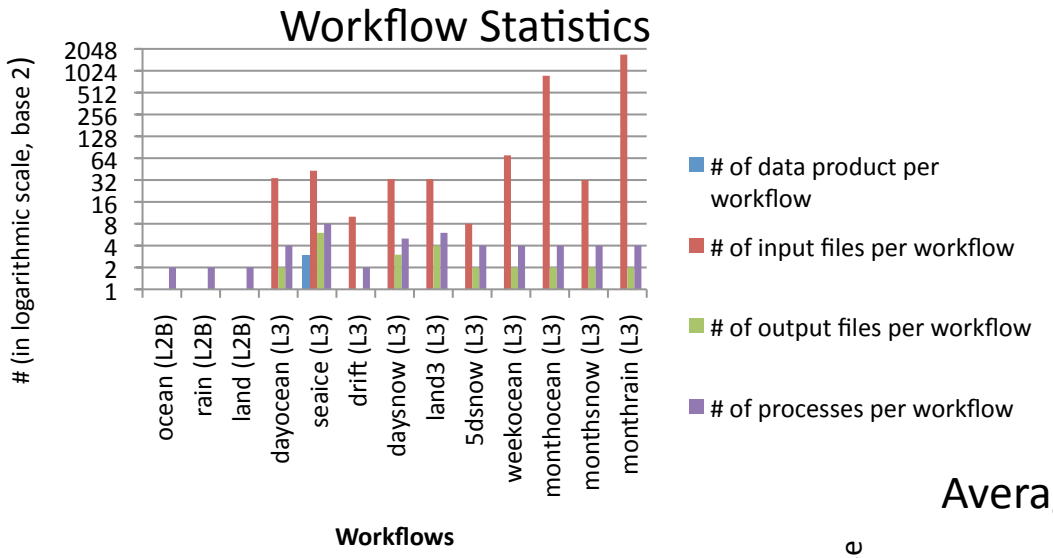
Provenance capture in Polar Orbiting Satellite Imagery Ingest Process

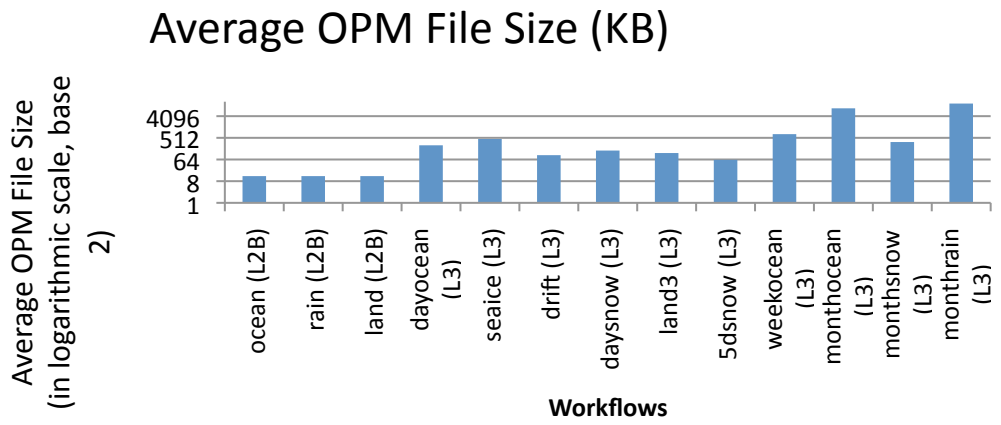
NASA AMSR-E imagery ingest processing schedule.

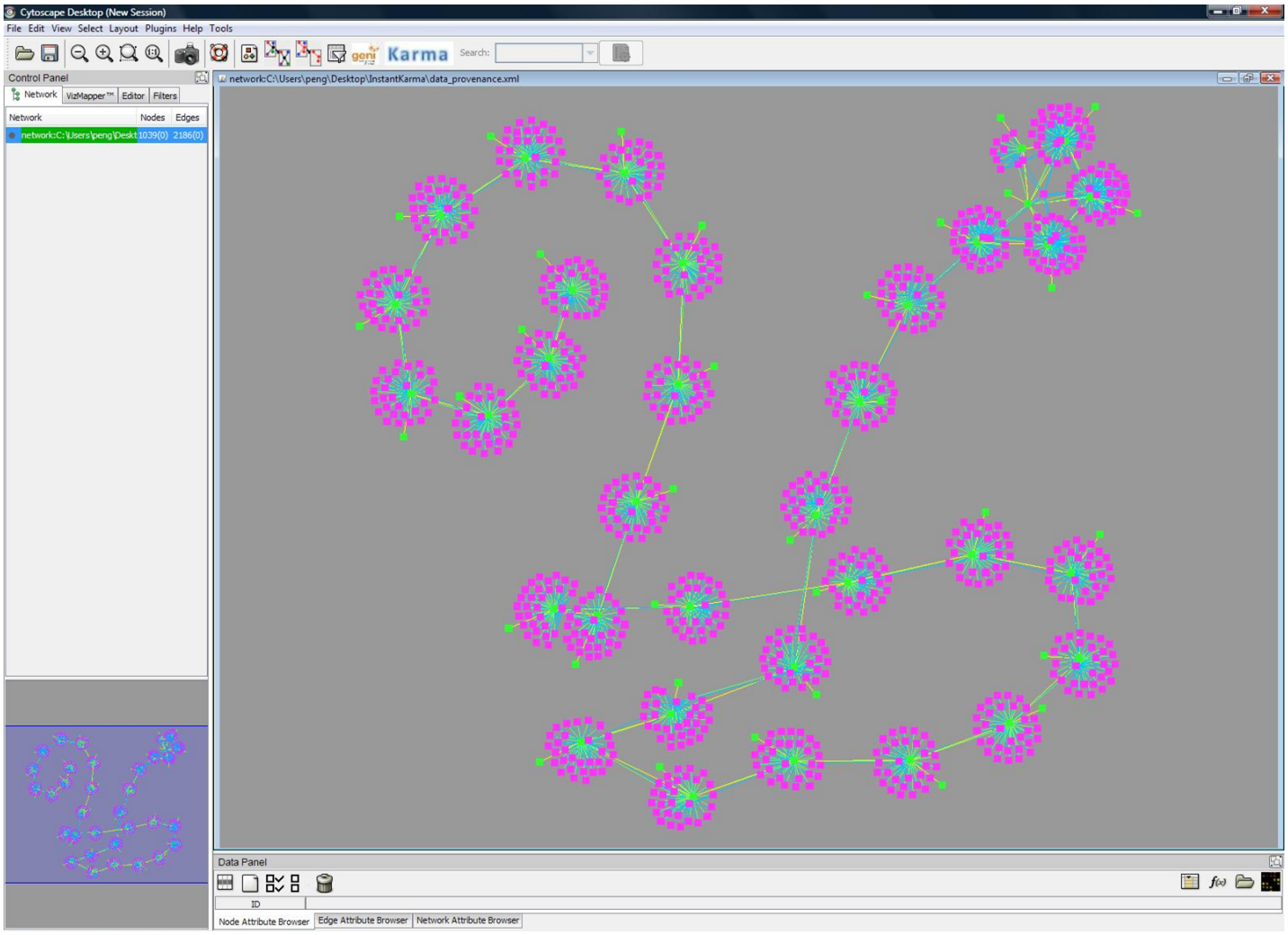
We captured provenance for all products for 1 month



Effort: reprocess all data for Sep 2011

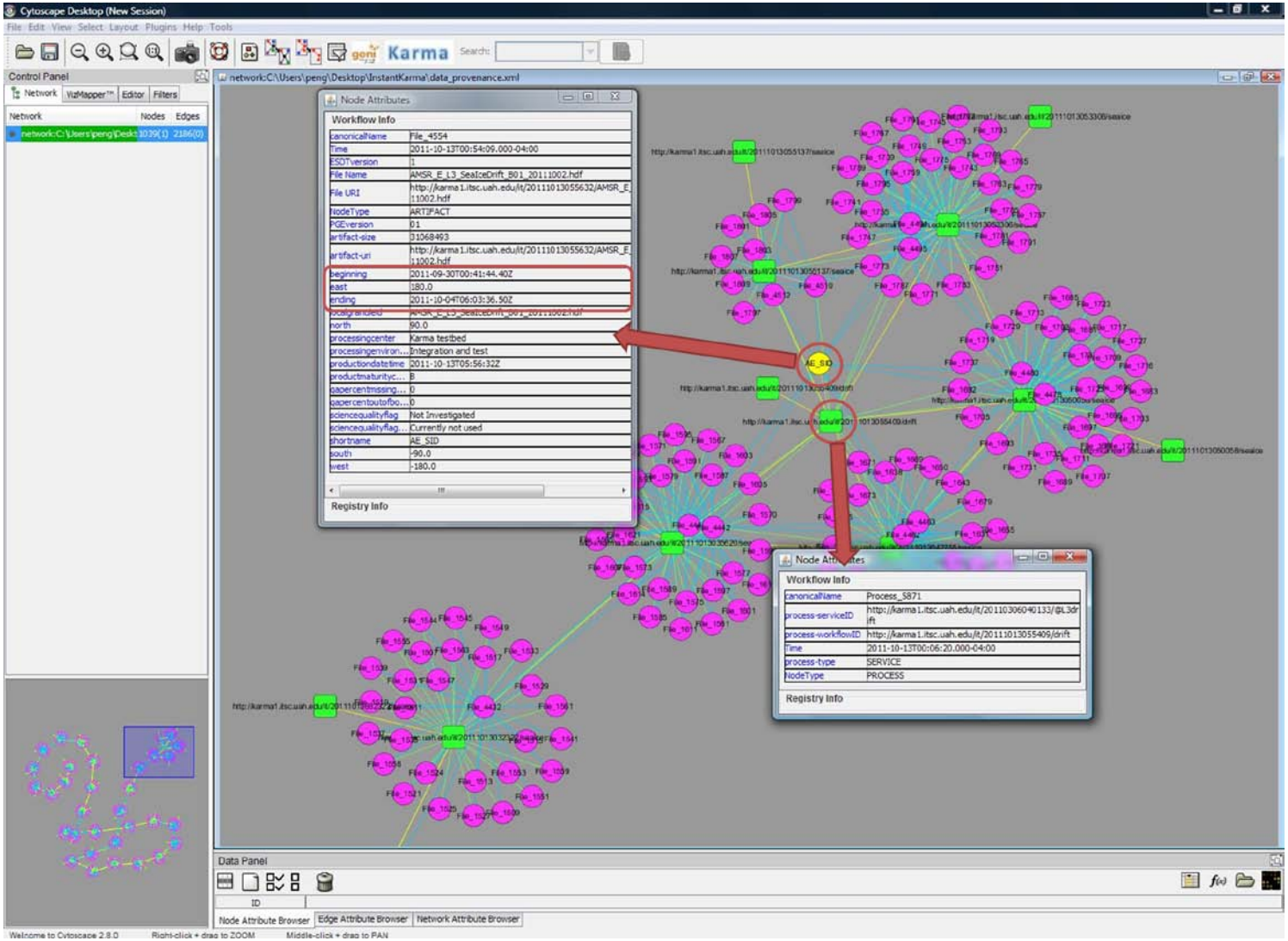




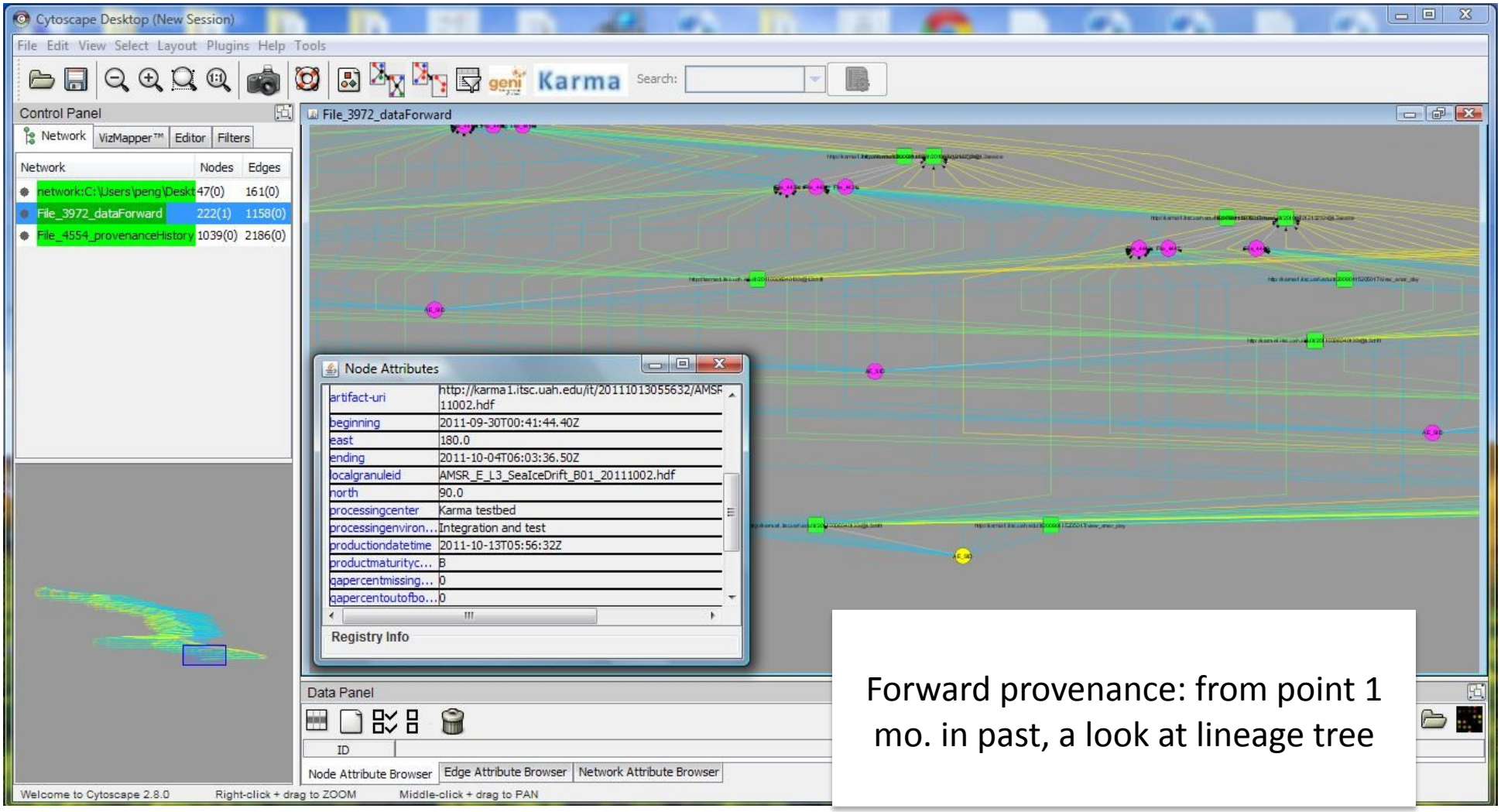


Provenance over linked workflows:

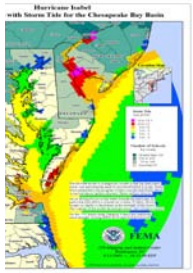
1 month of processing. Data products are related by moving mask file (temporal window; sea ice “stencil”) that is used.



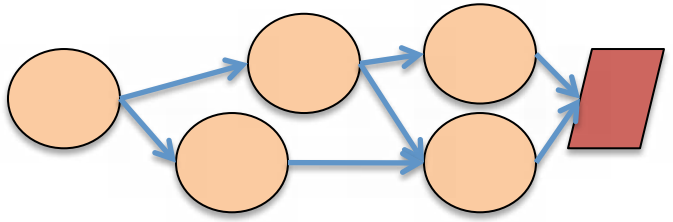
Last product generated in 1 month of processing (shown as yellow oval with red circle around)



Forward provenance: from point 1 mo. in past, a look at lineage tree

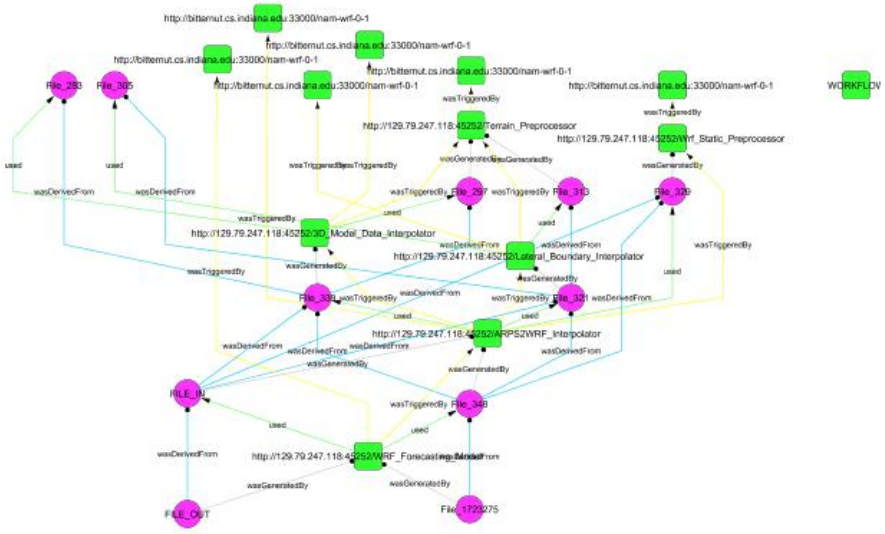



Workflow; steps to build

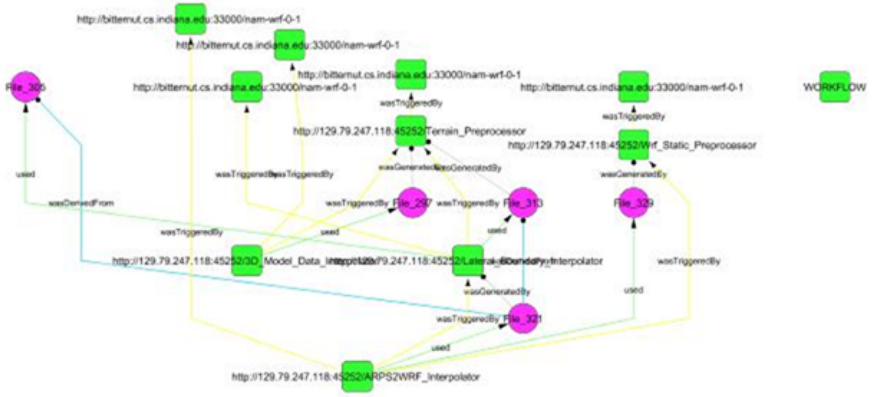


- User logs into web portal,
- Through user interface constructs workflow as directed graph of tasks executed in sequence. Edges are flows of data.
- Workflow (graph) handed off to scheduler that executes each task on cluster or cloud

Visualization of provenance of weather forecast workflow without and with failures



a



b

PRAGM

GENI Experiment: WiMAX DDoS

- We capture provenance of “DoS Attacks Exploiting WiMAX System Parameters”, Clemson. Experiment uses 100 subscribers with varied configurations of 6 parameters. Current version runs on NS2.

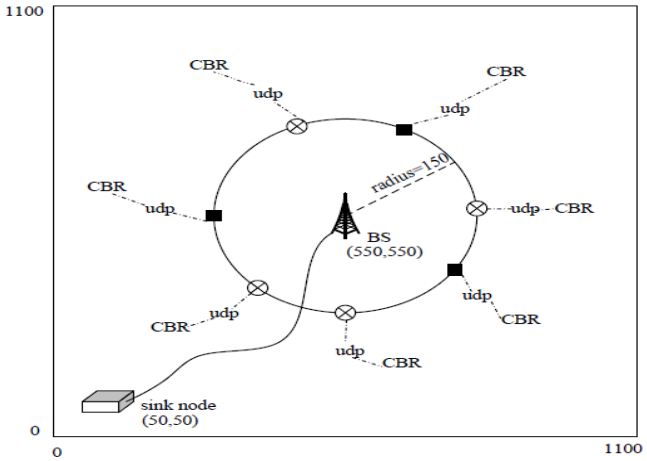


Fig.1: Network Topology

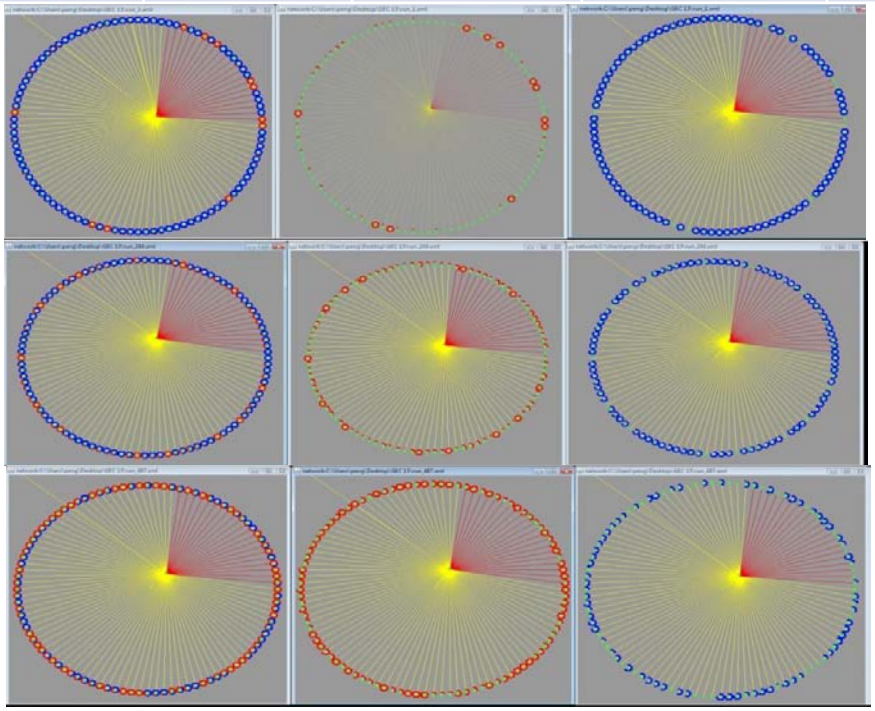
Table 1. Parameter Values

Parameter	Values		
	Treatment 1	Treatment 2	Treatment 3
frame_duration	0.004	0.01	0.02
number_of_attackers/user	20/80	50/50	80/20
dos_backoff_start	1	3	5
dos_request_retry	2	6	10
bw_backoff_start	1	3	5
bw_request_retry	2	6	10

Provenance of WiMAX DDoS Experiment

- Provenance capture with NetKarma. NetKarma captures
  - provenance of packet movement, and
  - infers critical provenance about packets that were dropped, and by doing so is able to convey information about DDoS attacks through visualization
  - Improvement over earlier hand-worked ANOVA analysis.
- NetKarma's provenance filters and visualization extensions for Cytoscape enable side-by-side performance comparison of different experiment configurations. Visualizations show packets dropped and received. Visualization automatically adjusted for provenance volume (of total number of packets sent.)

Run id	Frame duration	number of attackers	attack backoff start	attack request retry	bw backoff start	bw request retry
1	0.004	20/80	1	2	1	2
244	0.01	20/80	1	2	1	2
487	0.02	20/80	1	2	1	2

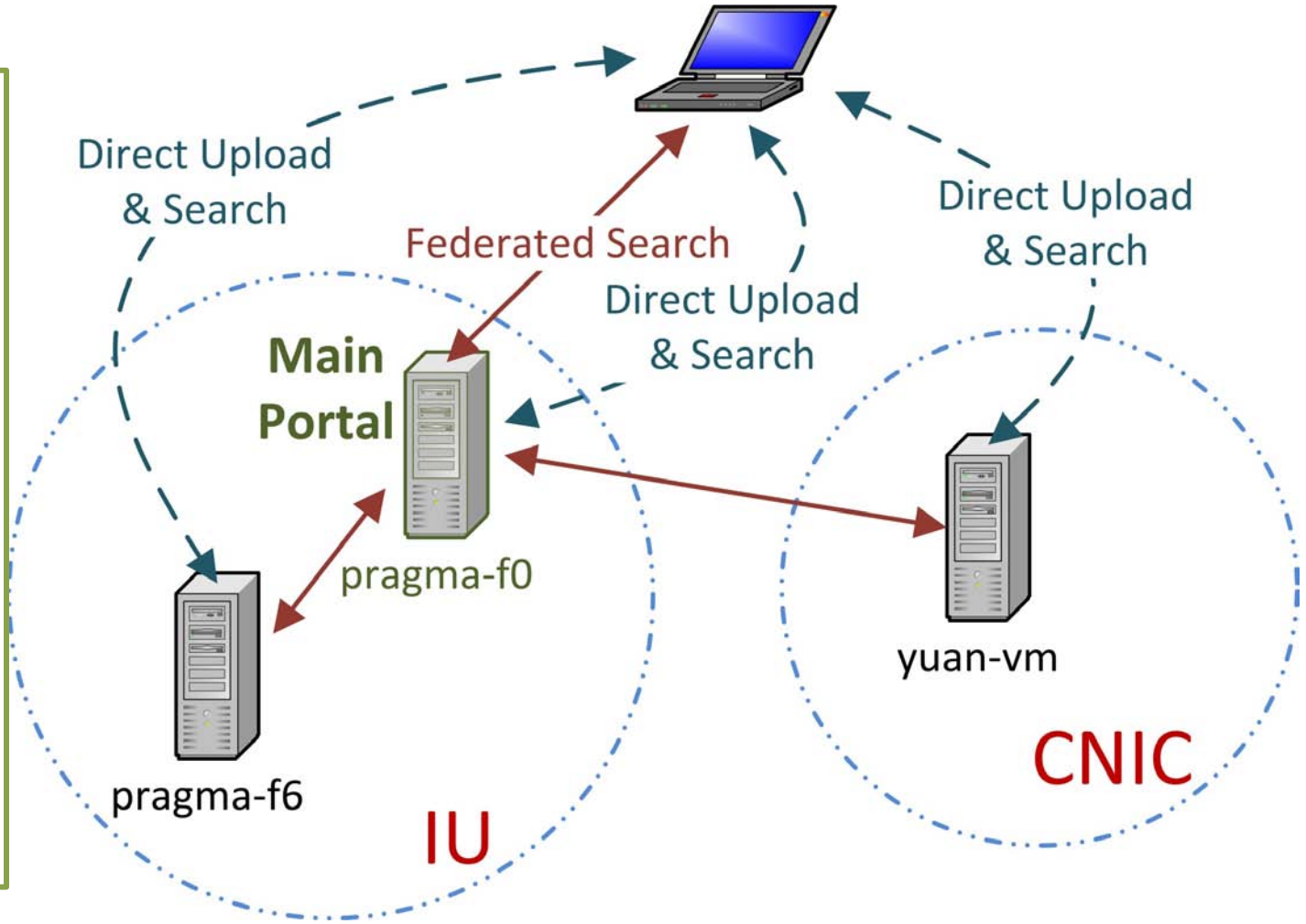


Dropped packets increases as frame duration increases from 0.01s to 0.02s

Shameless Plug for PRAGMA poster:

ESRI GeoPortal metadata servers used in federated mode on PRAGMA cloud.

Serves out weather forecast model data metadata (in FGDC) through main server or secondary server.



My interests are at intersection of metadata and semantics for geo kinds of e-Science.

Beth Plale
plale@indiana.edu