DASPOS Update

Mike Hildreth
representing the DASPOS project

Recent Work

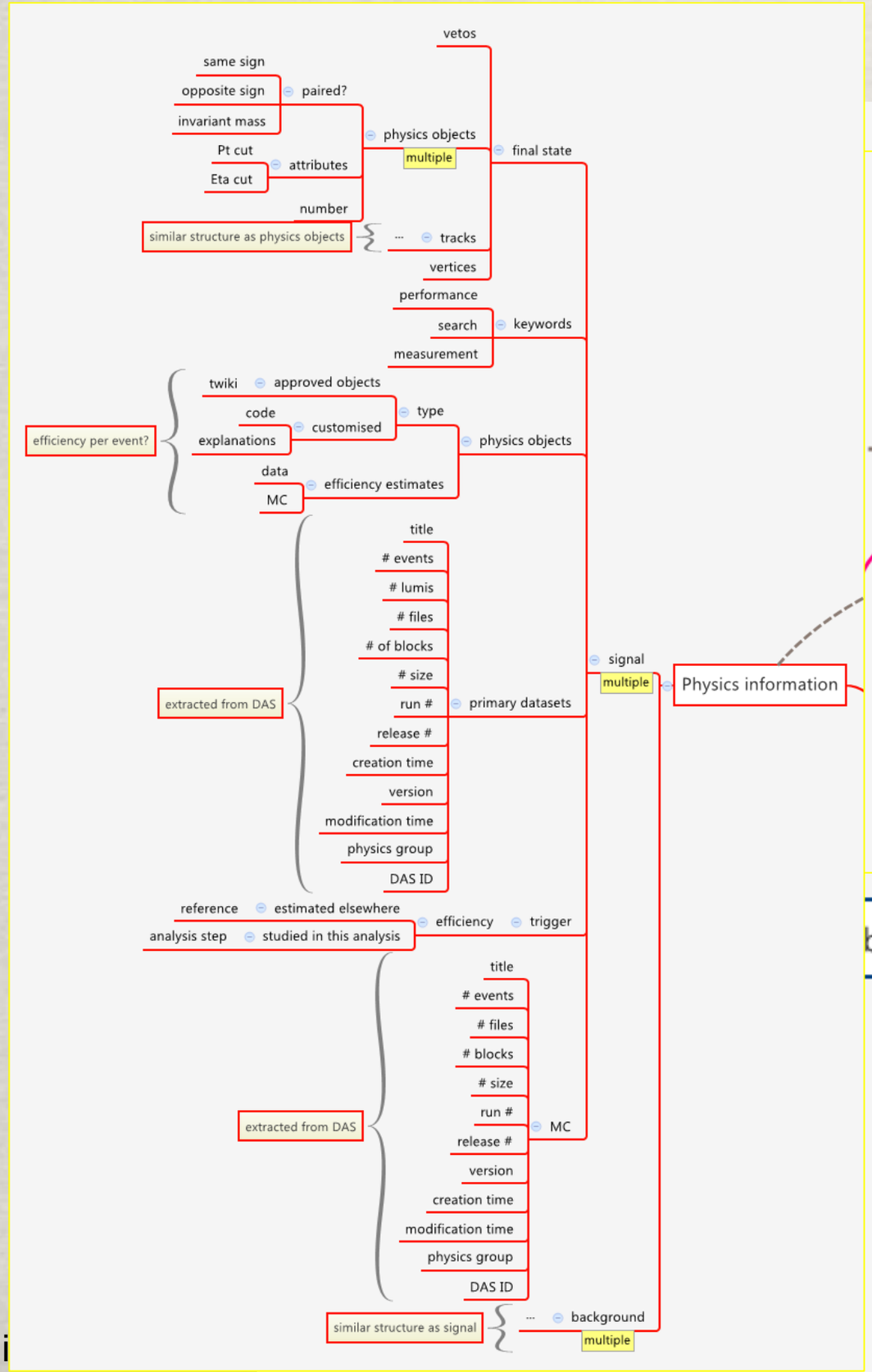
- HEP Data Model Workshop ("VoCamp15ND")
  - May 18-19, 2015, Notre Dame, IN
  - Participants from HEP, Libraries, & Ontology Community*

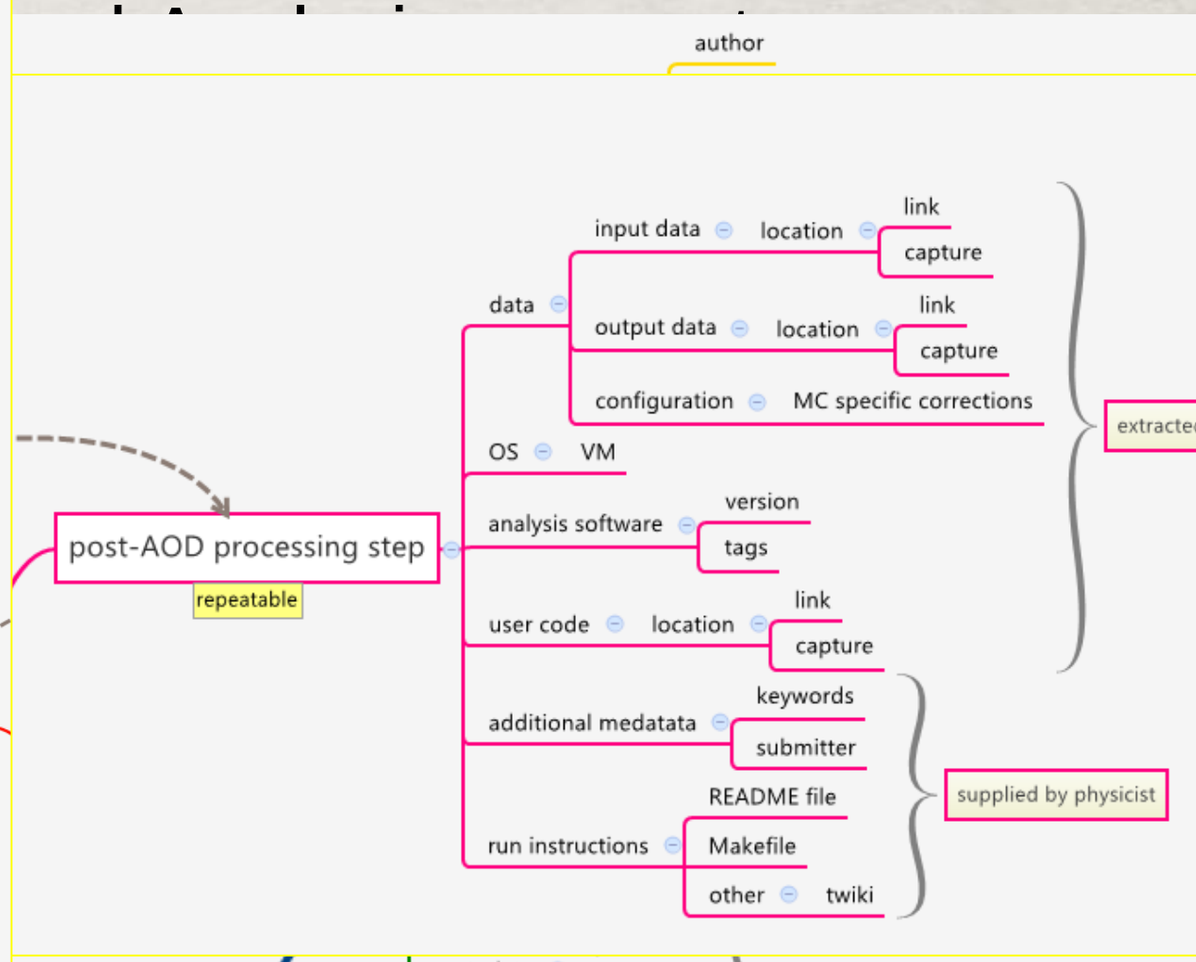
    *new collaborations for DASPOS
  - Define preliminary Data Models for CERN Analysis Portal
    - describe:
      - main high-level elements of an analysis
      - main research objects
      - main processing workflows and products
      - main outcomes of the research processw
    - potentially re-use bits of developed formal ontologies
      - PROV, Computational Observation Pattern, HEP Taxonomy, etc.

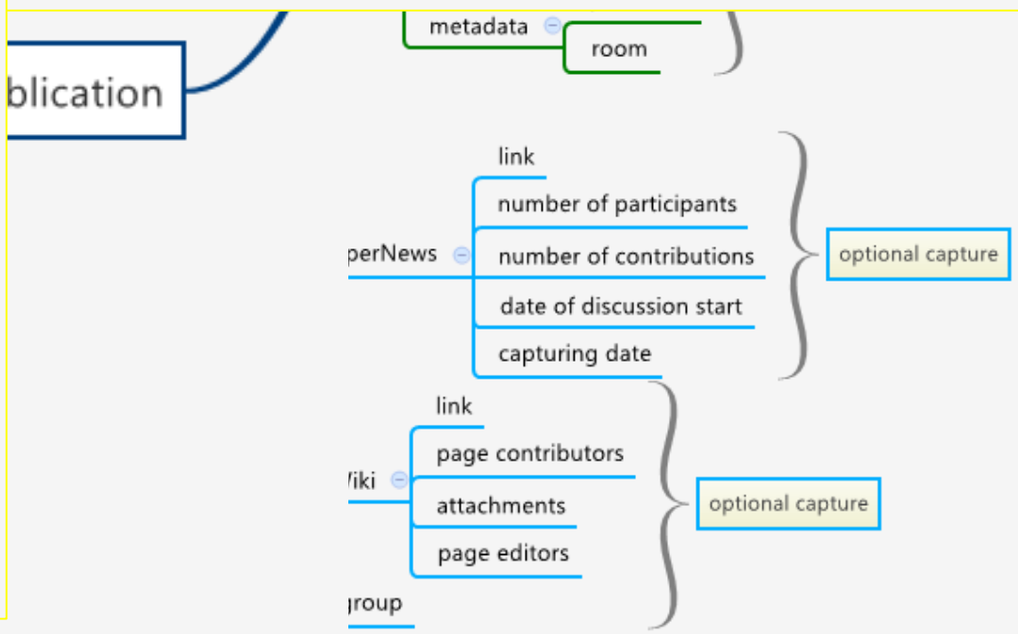
…small back story

- CERN Open Data Portal built using Marc21 as a description language/infrastructure
  - pushing boundaries of what is possible
    - fairly rigid xml-based data descriptions
  - new approach needed for analysis preservation
    - JSON-based format (e.g. JSON-DL) is both flexible and extensible
  - implementation of any new Data Models in an appropriate description language is required to set up the CAP

Starting points

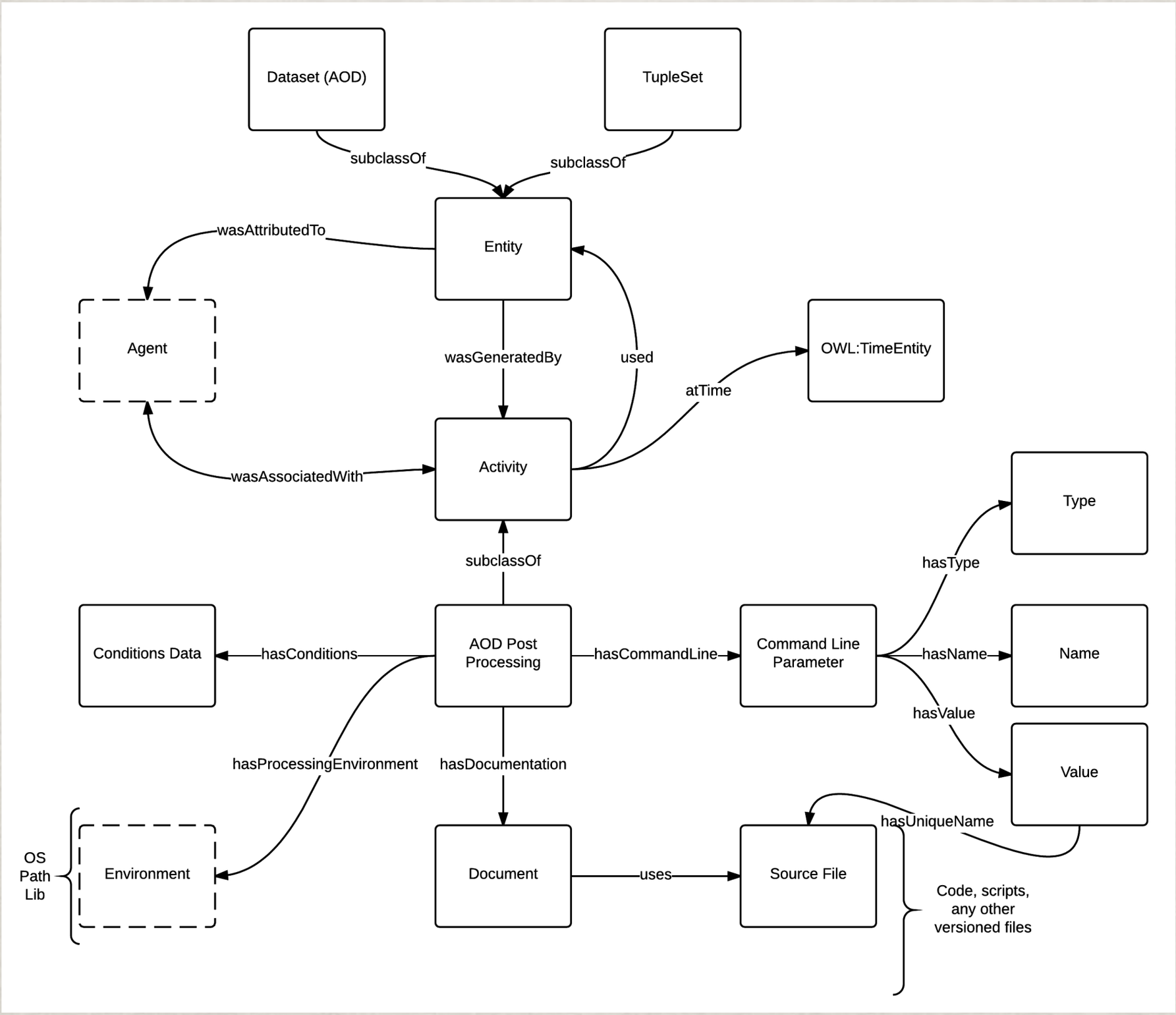




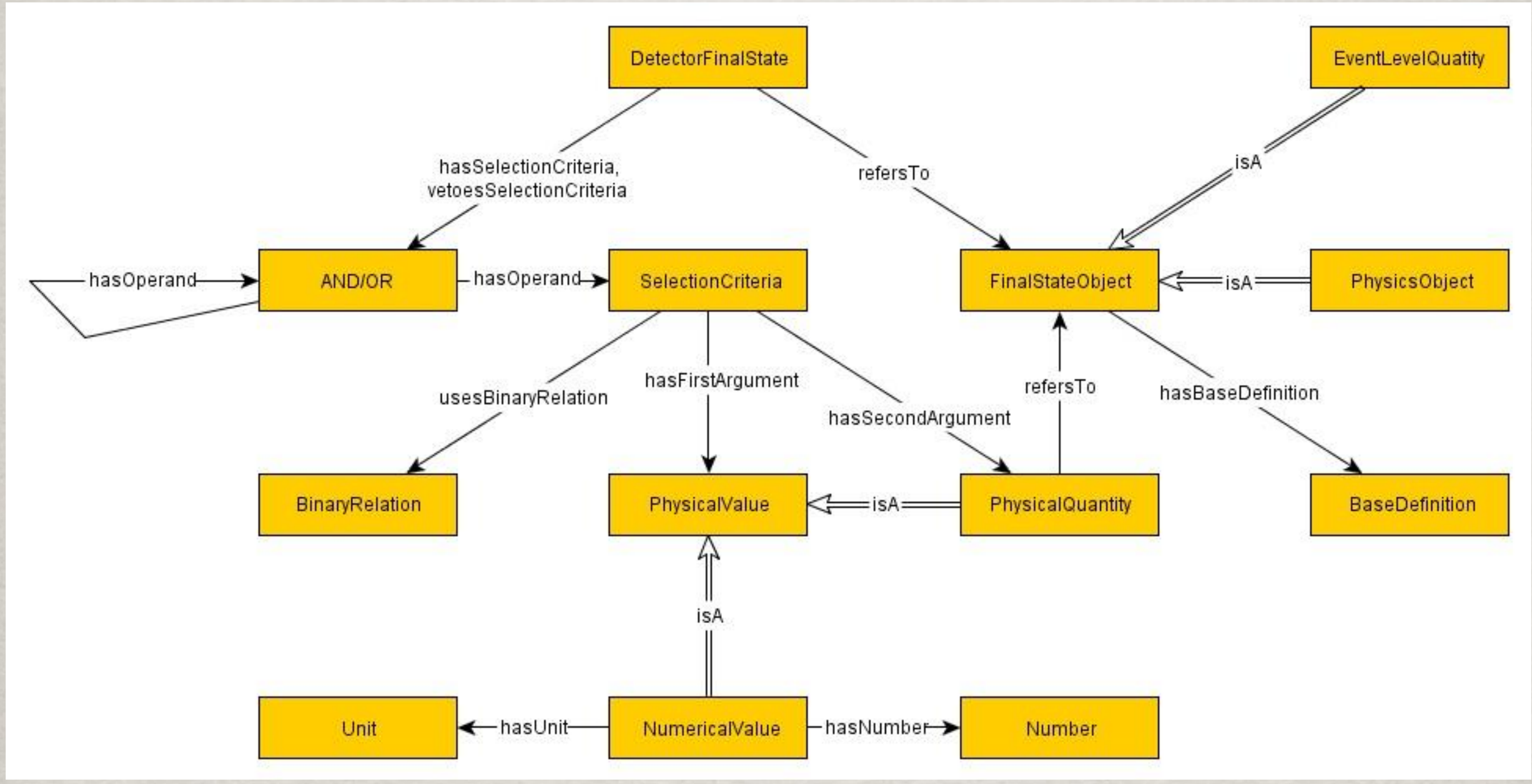


Mi

Workflow Description



Physics Description



Future Steps

- Implement Data Model(s) in appropriate JSON-based infrastructure (JSON-LD?)
  - CERN/DASPOS (ontologists)
- Population with test data
  - validation of description
  - allow simple internal queries
- In parallel: implementation of formal logic/ontological structure
  - enables full query/search/relational power of data model
  - DASPOS (ontologists)
- Build-out of Analysis Preservation Portal
  - deposit of “real” test cases
  - complicated/real query examples

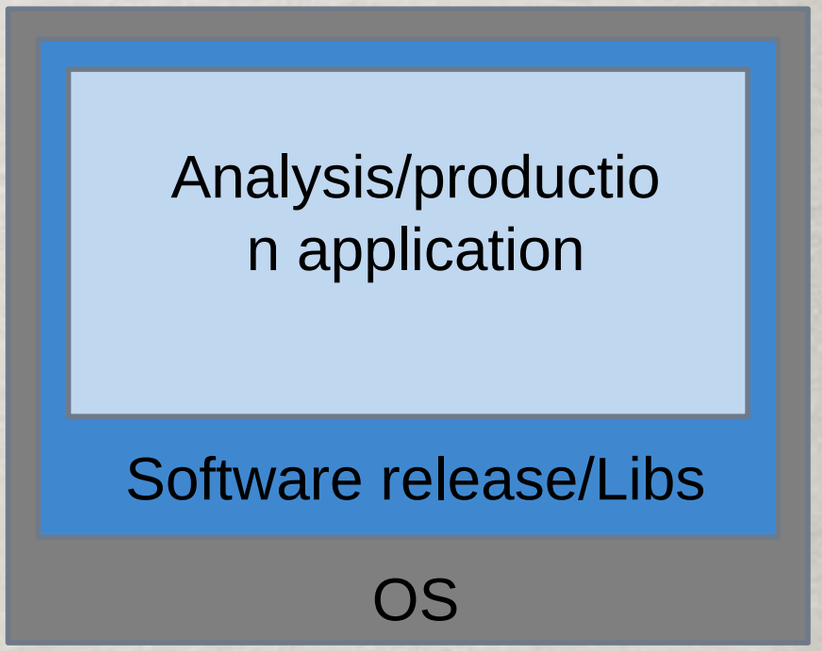
Other CAP-Supporting Activities

- Centered around capture and instantiation of preserved analyses
  - minimalist container capture
  - container description language (ND)
  - umbrella
    - environment “specification” and provisioning
    - pRUNe: workflow/provenance capture

- emphasis now on releasing v0.0 of all of these tools
  - partnership with CAP team to provide some back-end functionality

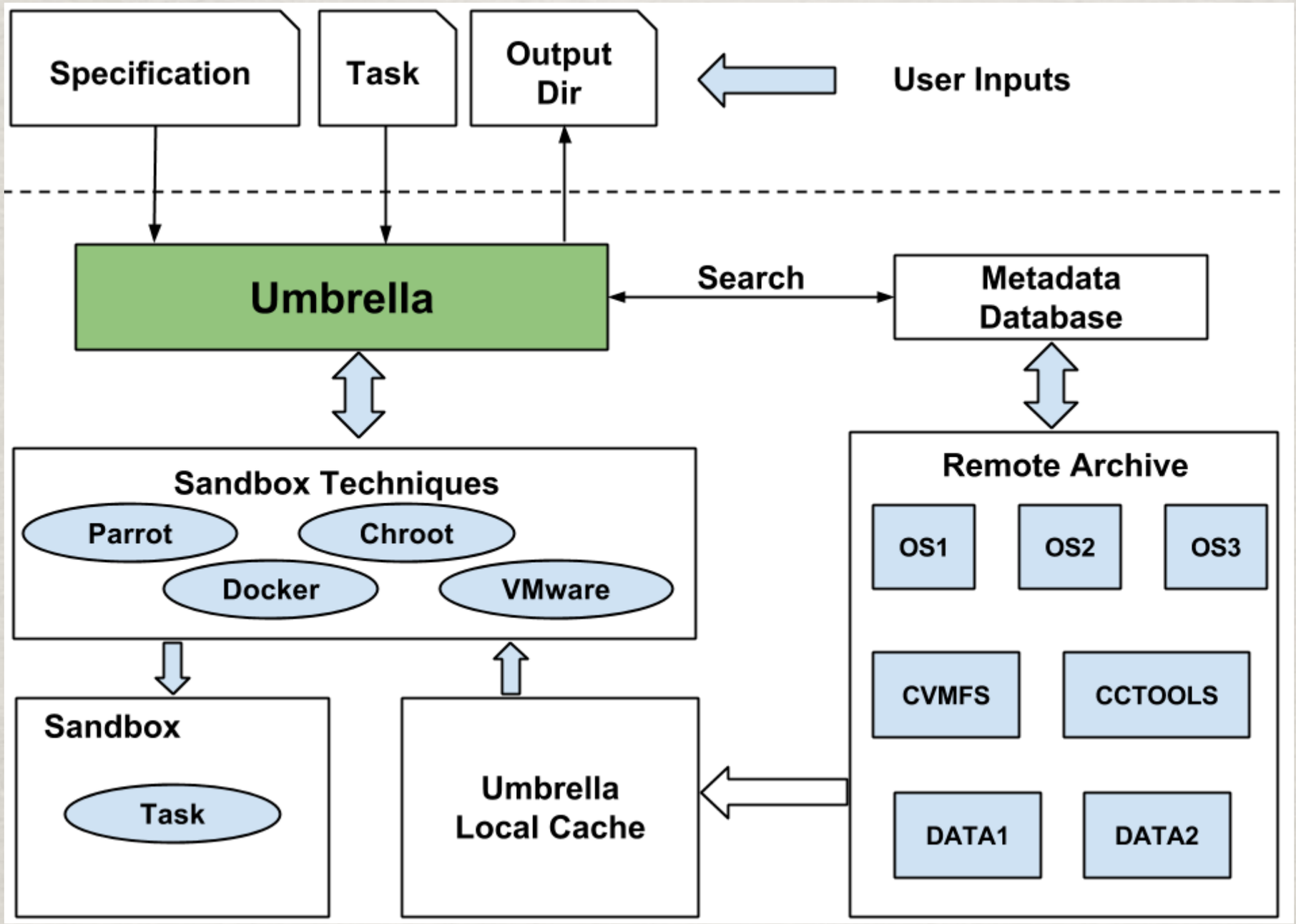
Containerization

- Minimum Containerized Unit for Analysis Capture? (UChicago)
  - to what extent can computing environment (and maybe even software releases) be separated from analysis workflows?
    - containers within containers
  - "Ease of capture" issues:
    - simpler for a user to bundle up local code rather than the entire software release/external libraries/etc.
    - also exploring CERNVM-FS in containers
  - Tools:
    - CDE, PTU, Container implementations
  - Plans:
    - stand up container-enabled back-end for some OSG services
      - test implementation of generic job execution with container submission



Umbrella

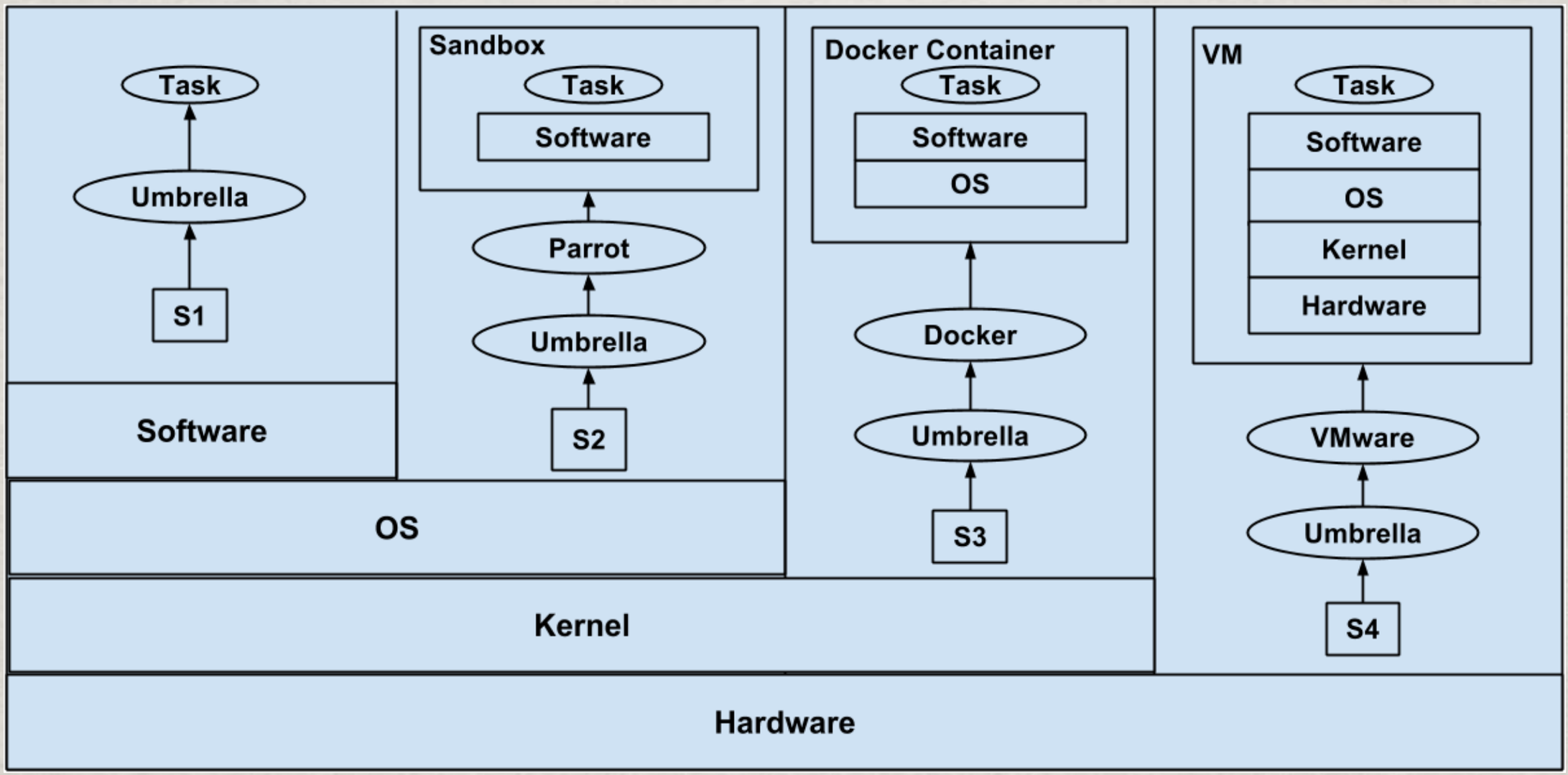
- Provisioning Framework (ND)



Umbrella

- Provisioning Framework (ND)

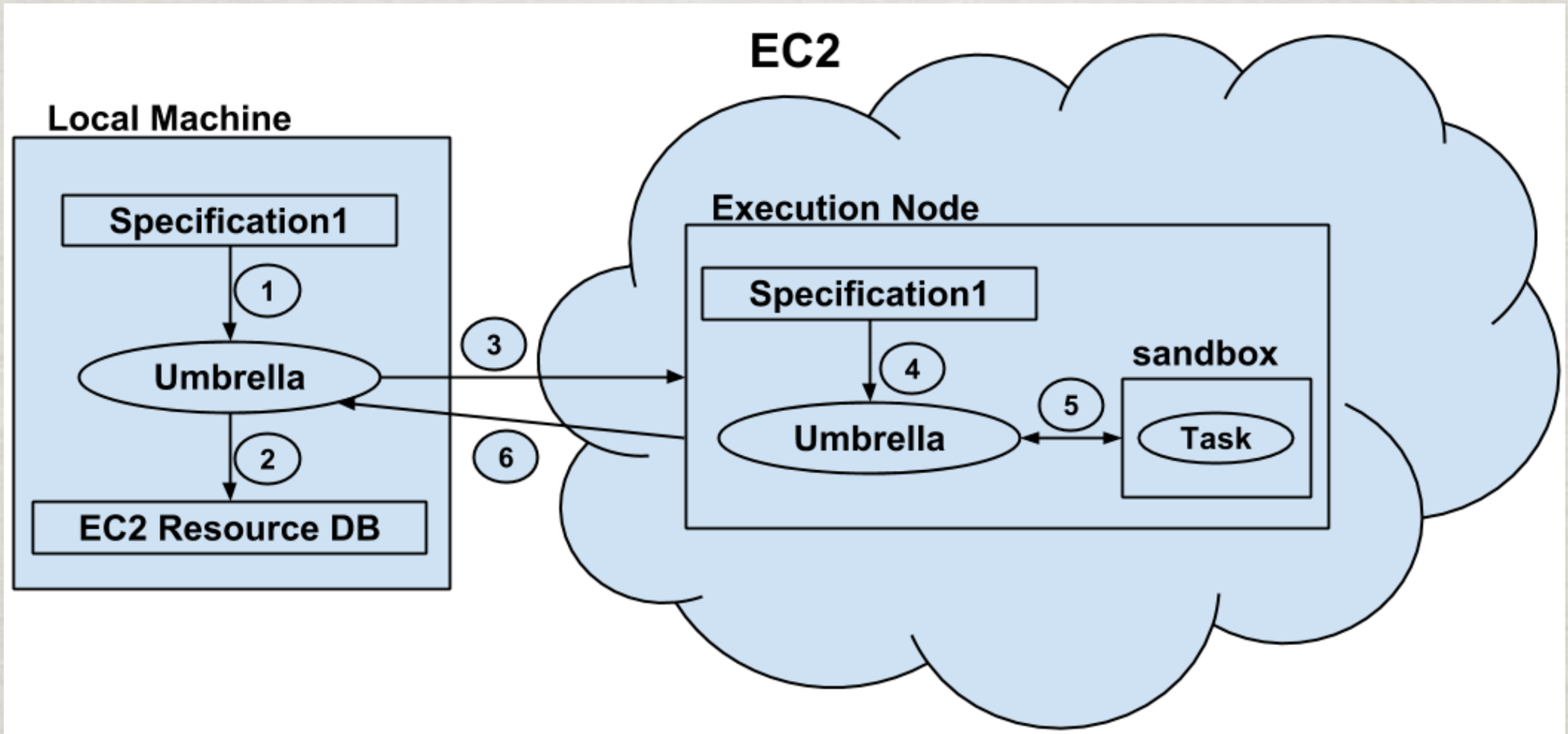
Varying Degrees of Virtualization



Umbrella

- Provisioning Framework (ND)

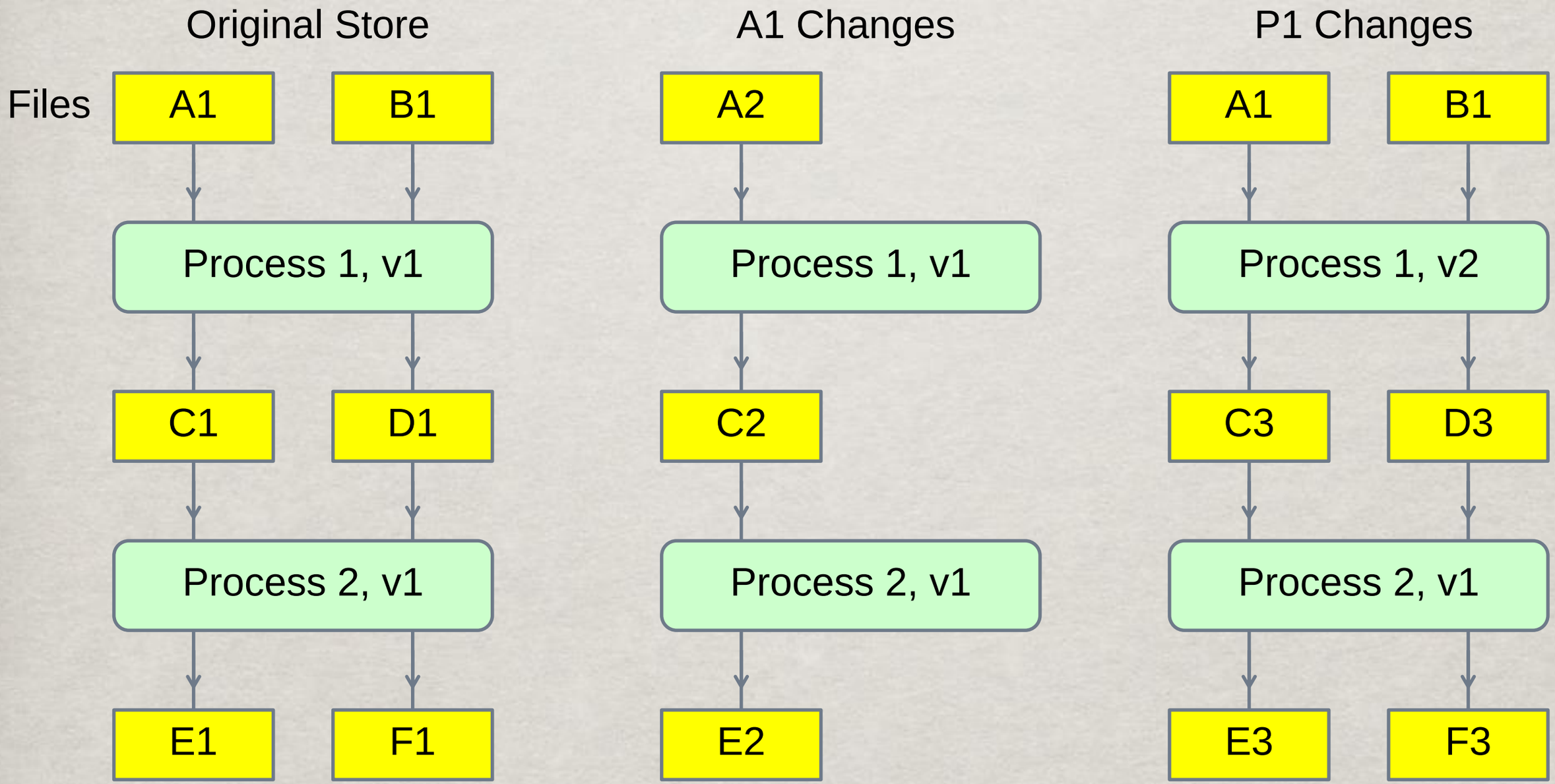
example: EC2 execution



- has been run in many different environments:
  - local Condor, OSG, local OpenStack, EC2

pRUNe

- Lightweight Workflow Provenance Capture (ND)



{E1: A1,P1v1,C1,P2v1}

{E2: A2,P1v1,C2,P2v1}

{E3: A1,P1v2,C3,P2v1}

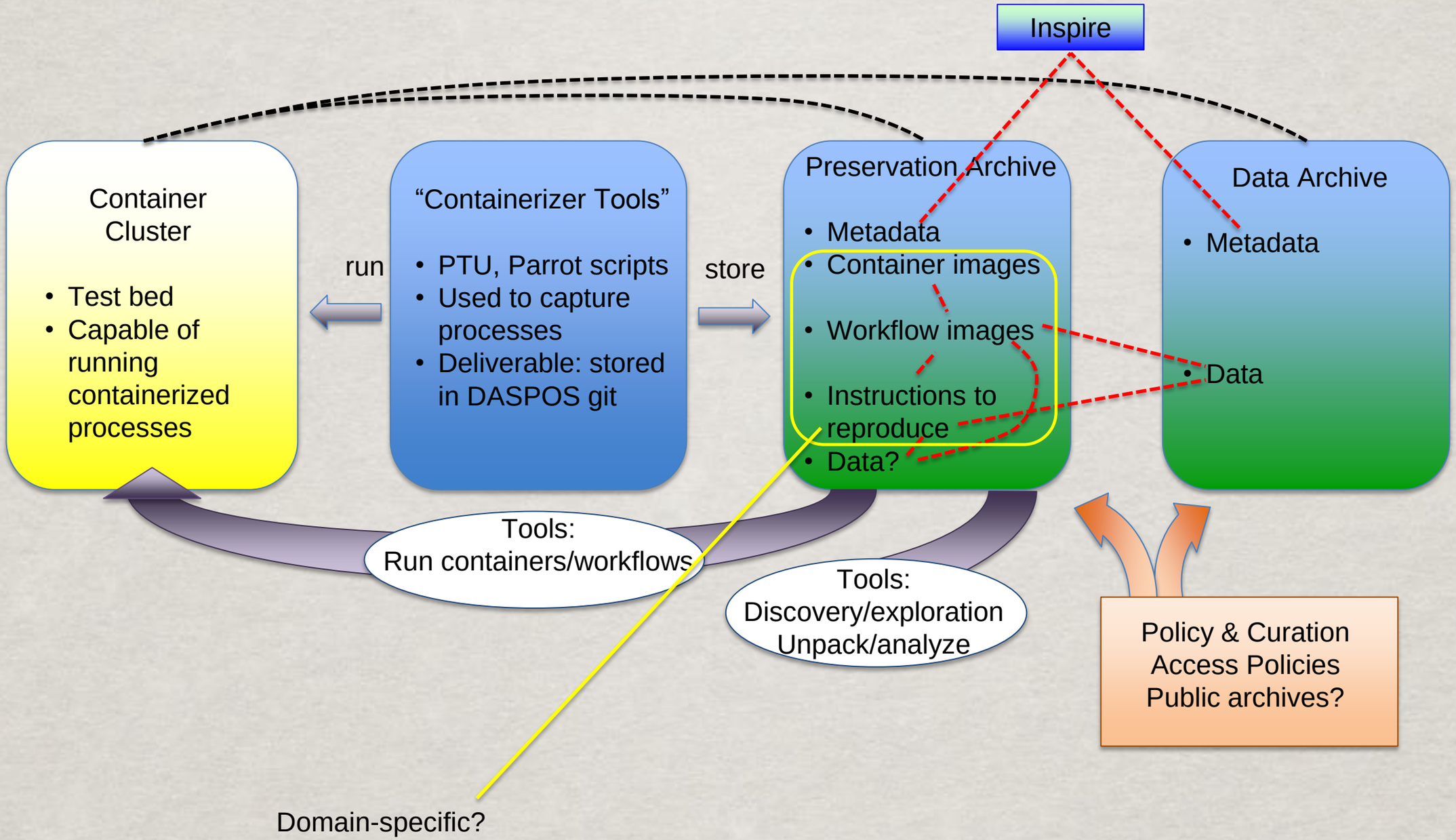
pRUNe

- Lightweight Workflow Provenance Capture (ND)
  - flexible framework to capture provenance while work is being done
    - assumes software environment is already specified
  - each object/process is captured in a database with a unique identifier
    - Building blocks: files and processes can be accessed and re-used
  - possible to automatically re-generate files later in processing chain with new versions of input files or processing steps
  - possible to export workflows and building blocks to new databases: portability

DASPOS Status

- Formal funding continues for one more year
- Will actively participate in CAP build-out
- Currently exploring future directions
  - NSF SSI (Sustainable Software)
  - OSG partnership

Possible Knowledge Preservation Architecture



DASPOS

- Data And Software Preservation for Open Science
  - multi-disciplinary effort recently funded by NSF
  - Notre Dame, Chicago, UIUC, Washington, Nebraska, NYU, (Fermilab, BNL)
- Links HEP effort (DPHEP+experiments) to Biology, Astrophysics, Digital Curation
  - includes physicists, digital librarians, computer scientists
  - aim to achieve some commonality across disciplines in
    - meta-data descriptions of archived data
      - What's in the data, how can it be used?
    - computational description (ontology development)
      - how was the data processed?
      - can computation replication be automated?
    - impact of access policies on preservation infrastructure