Provenance

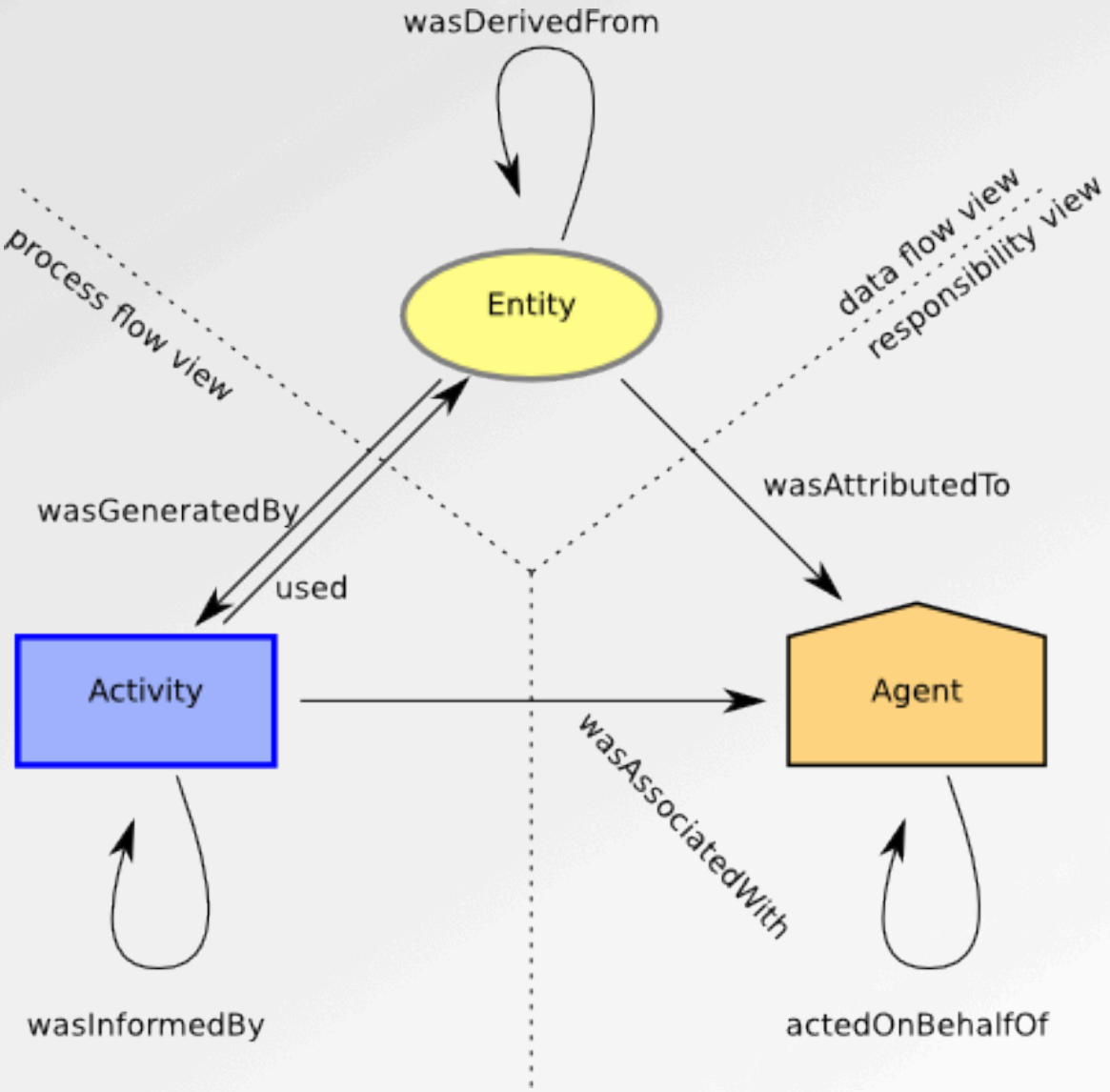
Provenance Views

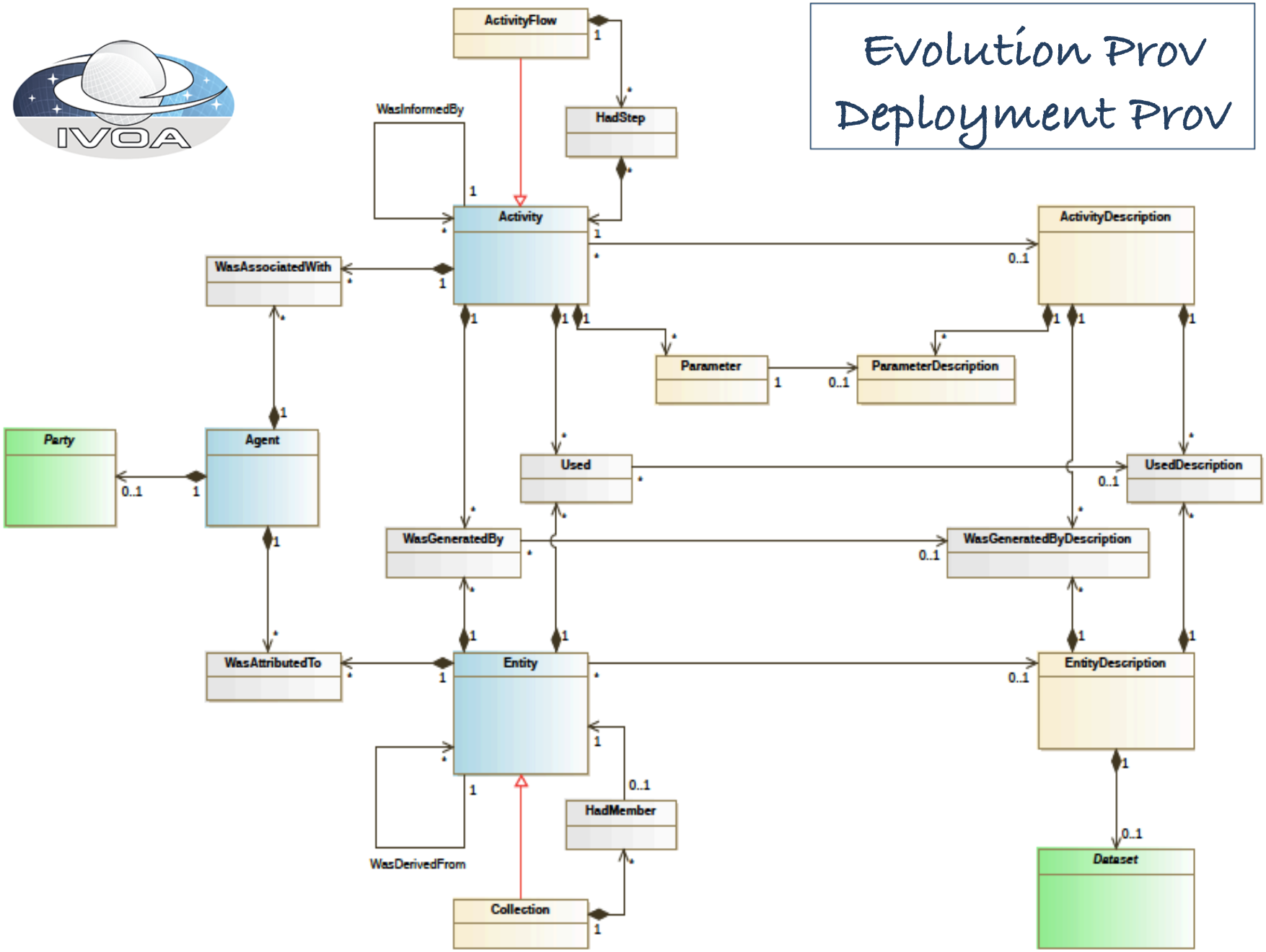
- Data
  - data quality and history inspection
  - mostly recorded in FITS HISTORY headers
  - exposed in a text file

- Experiment
  - definition provenance
    - structure / flowchart view of the **whole**
  - deployment provenance
    - execution environment
  - execution provenance
    - profiling and exec. logs
    - functions calls and variables
    - input / intermediate / output values

- Evolution
  → → → versioning

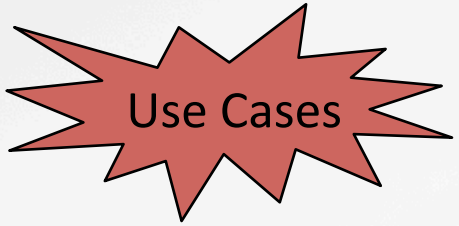
Provenance Views




Evolution Prov
Deployment Prov
ActivityFlow
HadStep
WasInformedBy
Activity
ActivityDescription
WasAssociatedWith
Parameter
ParameterDescription
Party
Agent
Used
UsedDescription
WasGeneratedBy
WasGeneratedByDescription
WasAttributedTo
Entity
EntityDescription
HadMember
WasDerivedFrom
Collection
Dataset

Application/User Levels in CTA

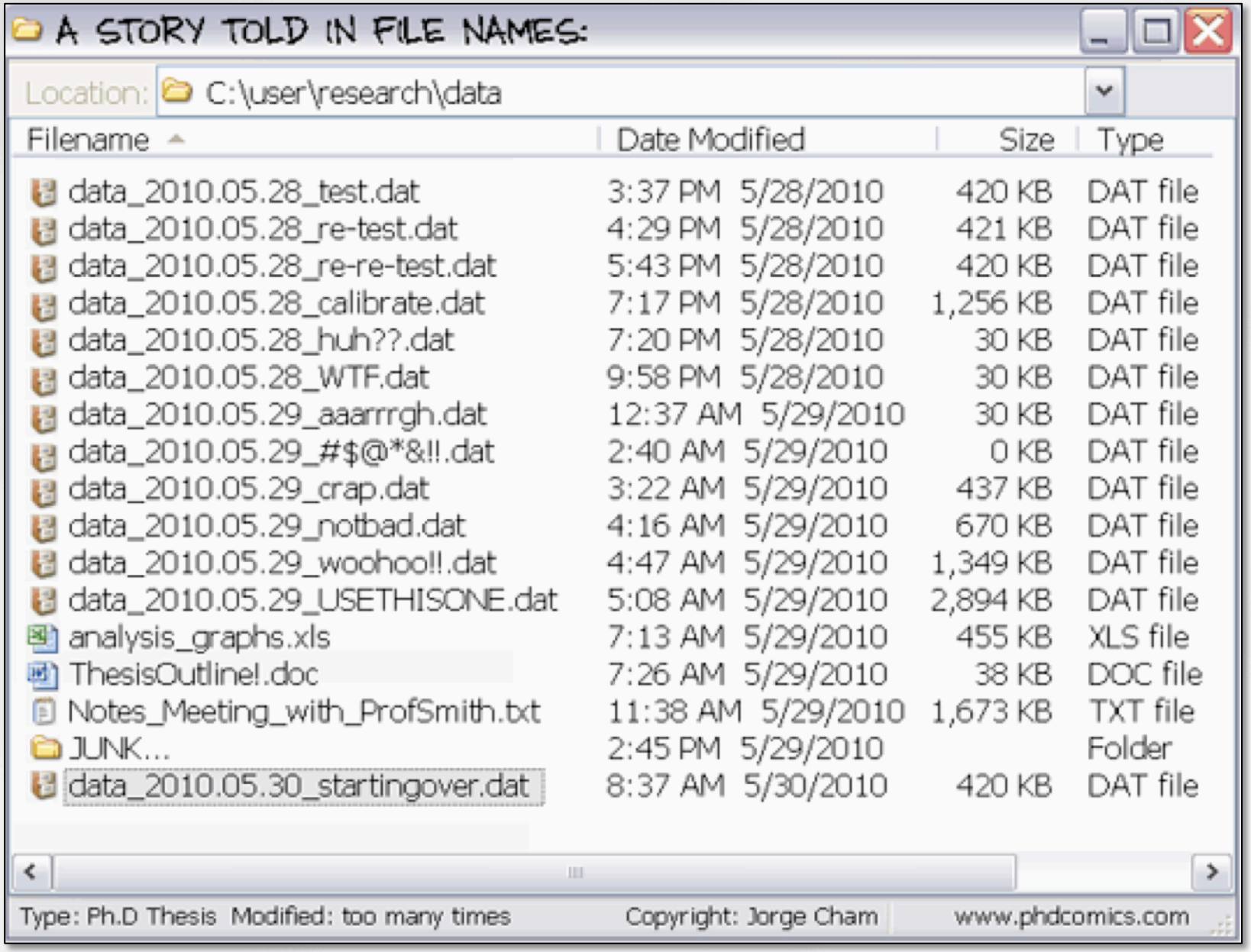
- Simulations
  - execution/evolution provenance
- Pipelines on raw data
  - execution provenance (exec. logs and profiling)
- Public DL3 → DL5 Archive
  - data provenance



- User Desktop (non-controlled environment)
  - definition provenance
    - structure / flowchart view of the experiment
  - deployment provenance
    - environment
  - evolution provenance
    - versioning tools

Provenance Capture in the Local Desktop

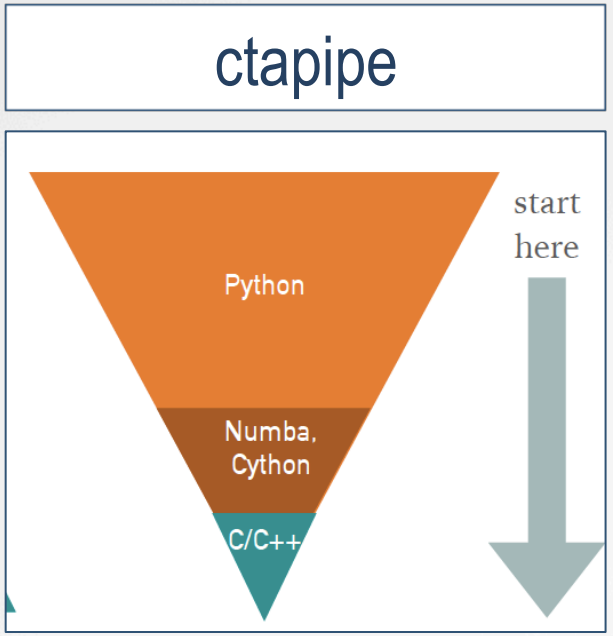
Provenance Capture in the Local Desktop



Provenance in script-based methodology

Scripts **orchestrate** analysis and **connect** data and tools

Python scripts as a glue



Challenges

- encode control/loops
- level of granularity
- non-controlled environment

Lesson learned

- prov. capture /inspection /analysis MUST be:
  - non-intrusive
  - user-friendly

noWorkflow Tool

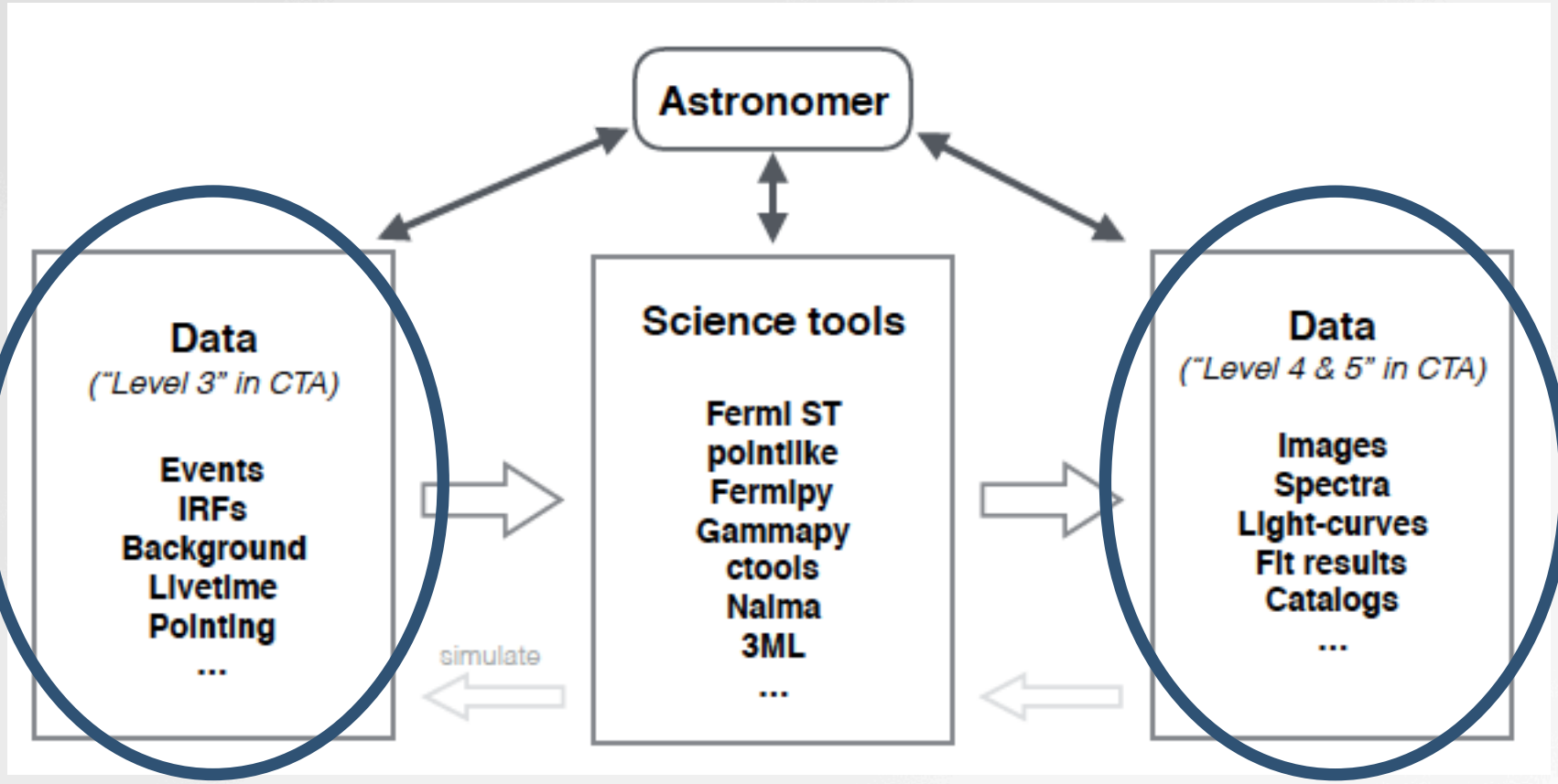
Captures process provenance for a **data analysis** working **methodology** based on **python scripts** and trial/error **exploration runs**.

- Provenance storage: SQLite DB + File System
- Provenance sharing of a local working session: `.noworkflow` folder
- Jupyter Notebooks support

Provenance capture of what happens inside a Python script

- Definition Provenance
  - Abstract Syntax Tree Analysis (code parsing / heuristics)
- Deployment Provenance
  - Python modules: `os, socket, modulefinder,...`
- Execution Provenance
  - Profiling and reflection (reimplementation of I/O functions)

Provenance in CTA Public Archive



- Pack of data files in a working session
- Connection with Local Desktop Provenance?

Querying with Provenance

Input Dimensions

- Identifier (entity/activity/agent)
- Focus
  - Data progenitors
  - Processes involved
  - Agents responsibility
  - Versioning
- Representation
  - Graph
  - Prov. File/VOTable
  - Whole Pack with Data Products
- Time direction (back/forward)
- Granularity
- Parameter vs. Language based queries

Exposing Provenance

Response Dimensions

- Representation
  - Graph
  - Prov. File/VOTable
  - Whole Pack with Data Products
- Analysis and Inspection
  - Graph
  - Diff-based
  - Browseable and granularity (Links to prov. services in Datalink)
- Pack structure → Reproducibility
  - Prov. file as the descriptor of a pack of interlinked files
  - Data, documentation, scripts,...

Storing Provenance in Archive

- Provenance is about **relationships**
- Despite their name RDBs are not well suited for relationships
- Fixed schema of RDBs do not adapt well to changes
- Relationships are first priority in noSQL/**Graph** databases