Provenance

WHERE
WHEN
WHY
SOLUTION
WHO
HOW
WHAT
BUSINESS PLAN
time
money
quality
USER
CLOUD
30%
70%
cost
quality
price
EXPENSES
TAX
PROFIT
INCOME

INSTITUTO DE ASTROFÍSICA DE ANDALUCÍA, IAA-CSIC

IAA – CSIC

Technical Experience

Radio astronomy

Jose Enrique Ruiz
@bultako

VO Archives - Modelling and Implementation

IVOA Contributions

- Note. Scientific Workflows in the VO
- REC. PDL Parameter Description Language
- Draft. N-Dimensional Cube Model

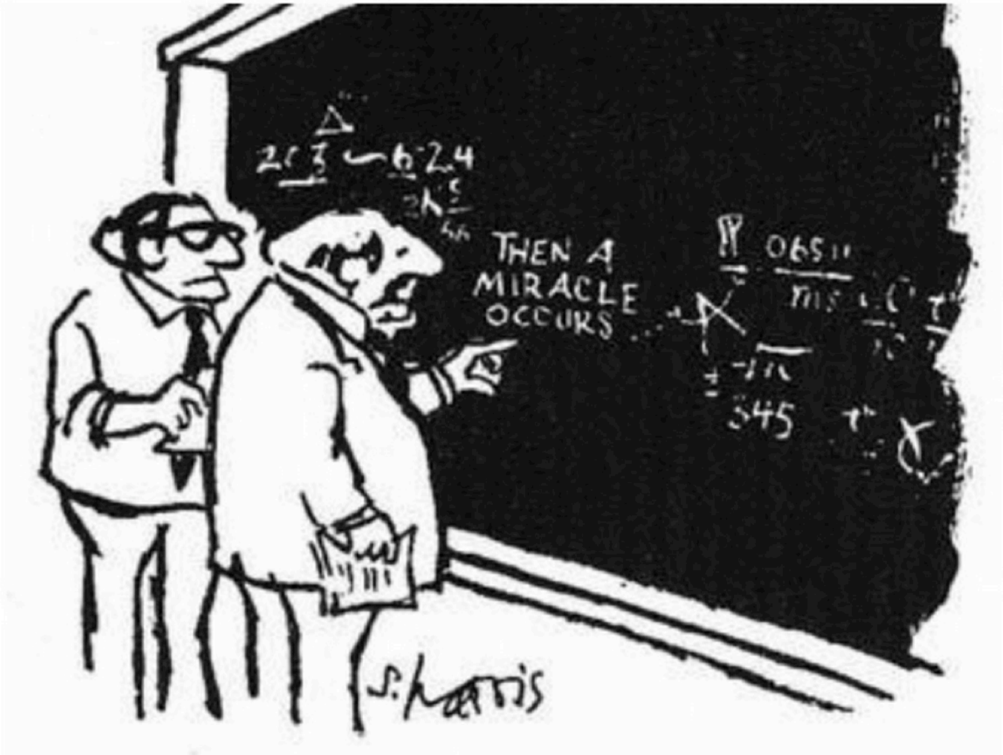
Software Development

- AstroTaverna – Building Workflows in the VO
- GUIPSY– Kinematic modelling for velocity datacubes of galaxie

Reproducibility / Credibility Crisis

"... up to 70% of research from academic labs **cannot be reproduced**, representing an enormous waste of money and effort."

- Elizabeth Iorns, Science Exchange



"I think you should be more explicit here in step two."

Digital Astronomy Cooking

Astronomy research lifecycle is **entirely digital**

- Observation proposals
- Data reduction pipelines
- Analysis of science ready data
- Catalogs of objects and data archives
- Publish process
  - Final data results
  - Experiment in Digital Libraries

    ADS/arXiv

Reproducible research is still not possible in a digital world

A normalized preservation of methodology is needed

Tools

A rich infrastructure of data is not efficiently used

Views

- Data Provenance
  - mostly recorded in FITS headers
  - data quality and history inspection
  - archive structured database modelling

- Process Provenance
  - definition provenance
    - structure / workflow view of the experiment
  - deployment provenance
    - execution environment
  - execution provenance
    - exec. log
    - functions calls, vars, input / intermediate / output values

- Evolution Provenance

Application Levels

- Low level pipelines raw data process
  - execution provenance (exec. log)
- Simulations
  - execution / evolution
- Public DL3 ->DL5 Archive
  - data provenance

- User Desktop
  - definition provenance
    - structure / workflow graph. view of the experiment
  - deployment provenance
    - environment
  - evolution provenance

Users

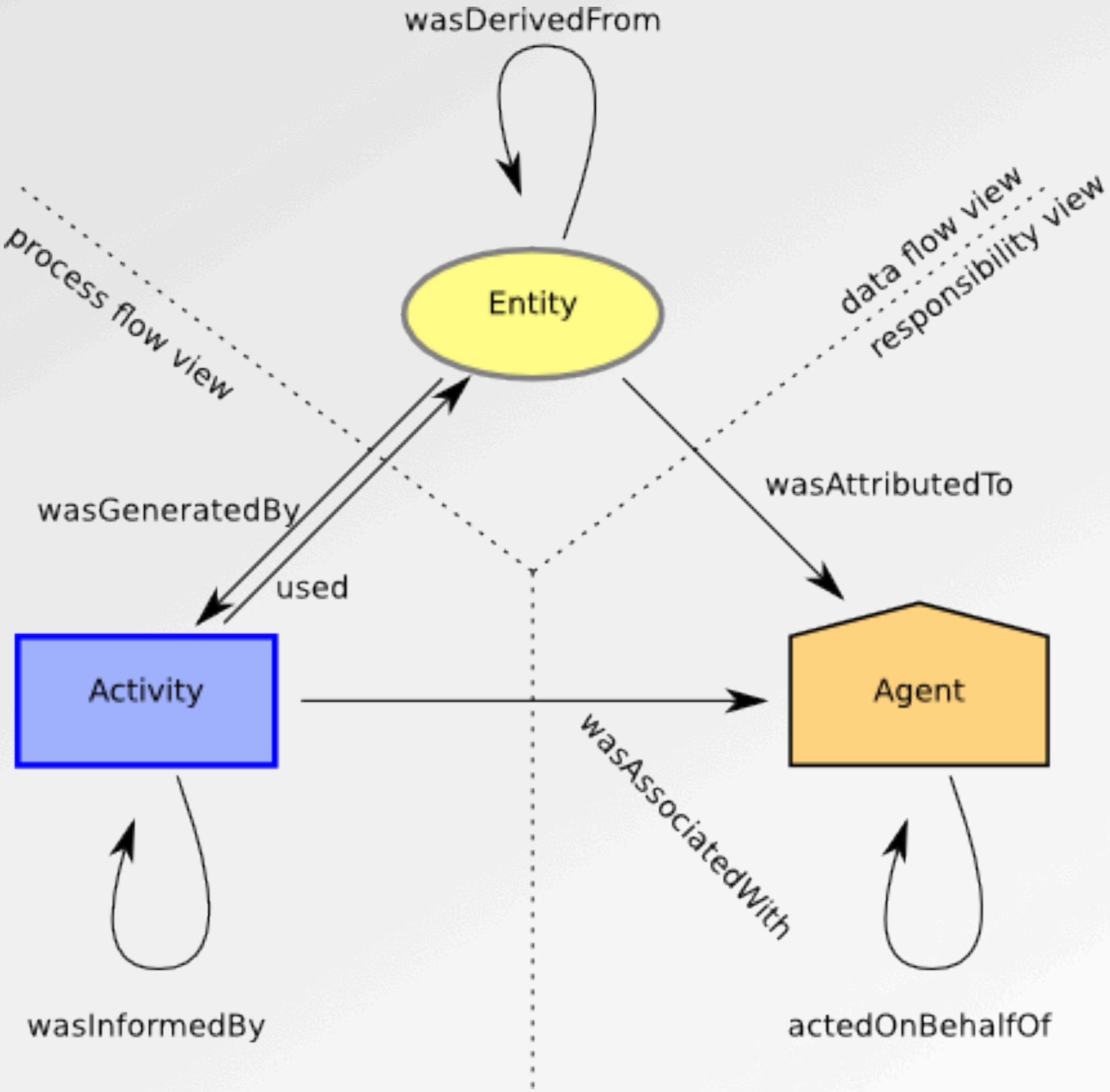
- Developers
  - Low level pipelines raw data process
    - execution
  - Simulations
    - execution / evolution
  - Analysis Tools
    - data / deployment / evolution

- Consumers
  - VHE Astronomers
    - data / deployment / evolution
  - Others
    - definition / deployment / evolution

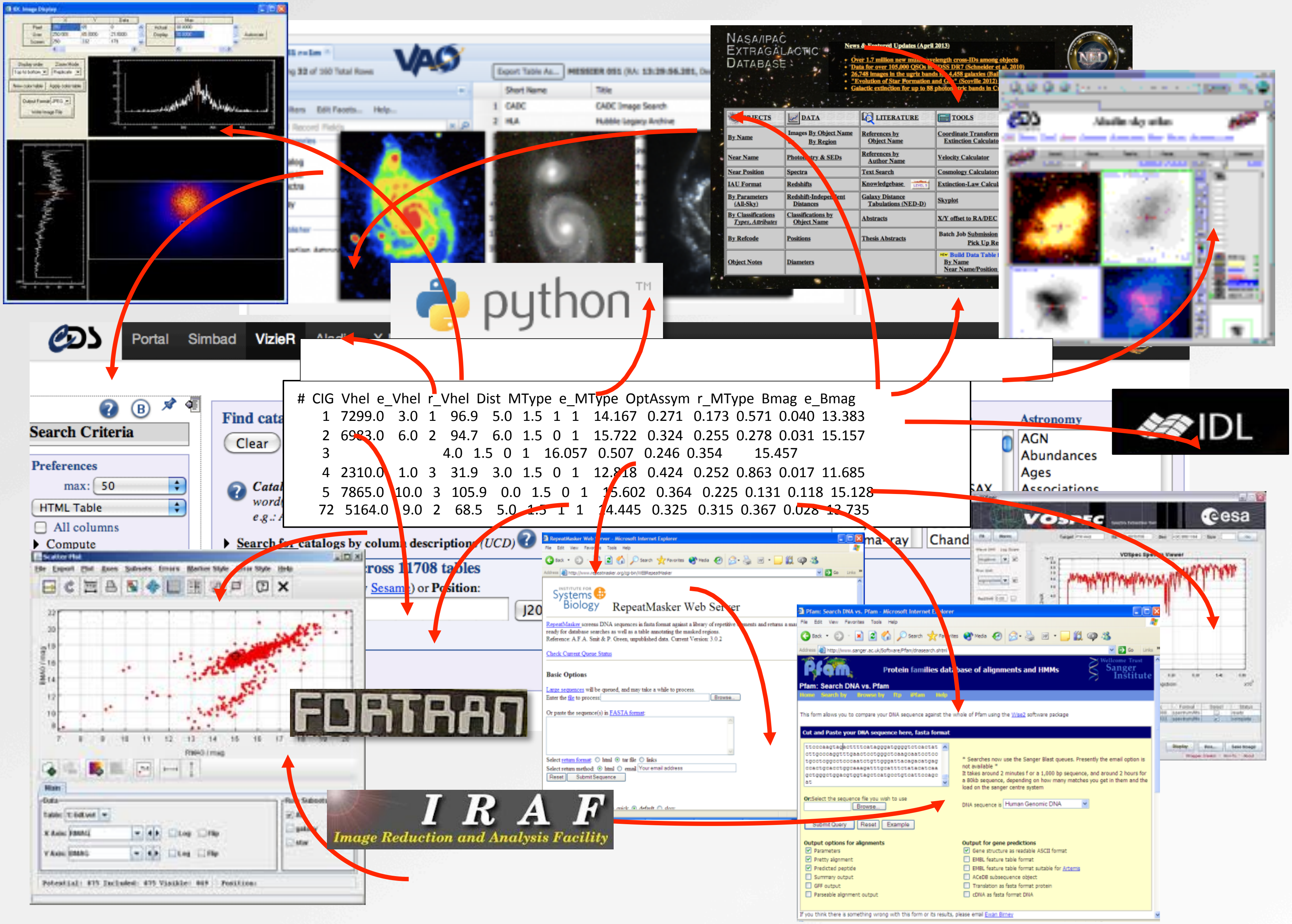
Use Cases

INSTITUTO DE ASTROFÍSICA DE ANDALUCÍA, IAA-CSIC

Structuring Provenance



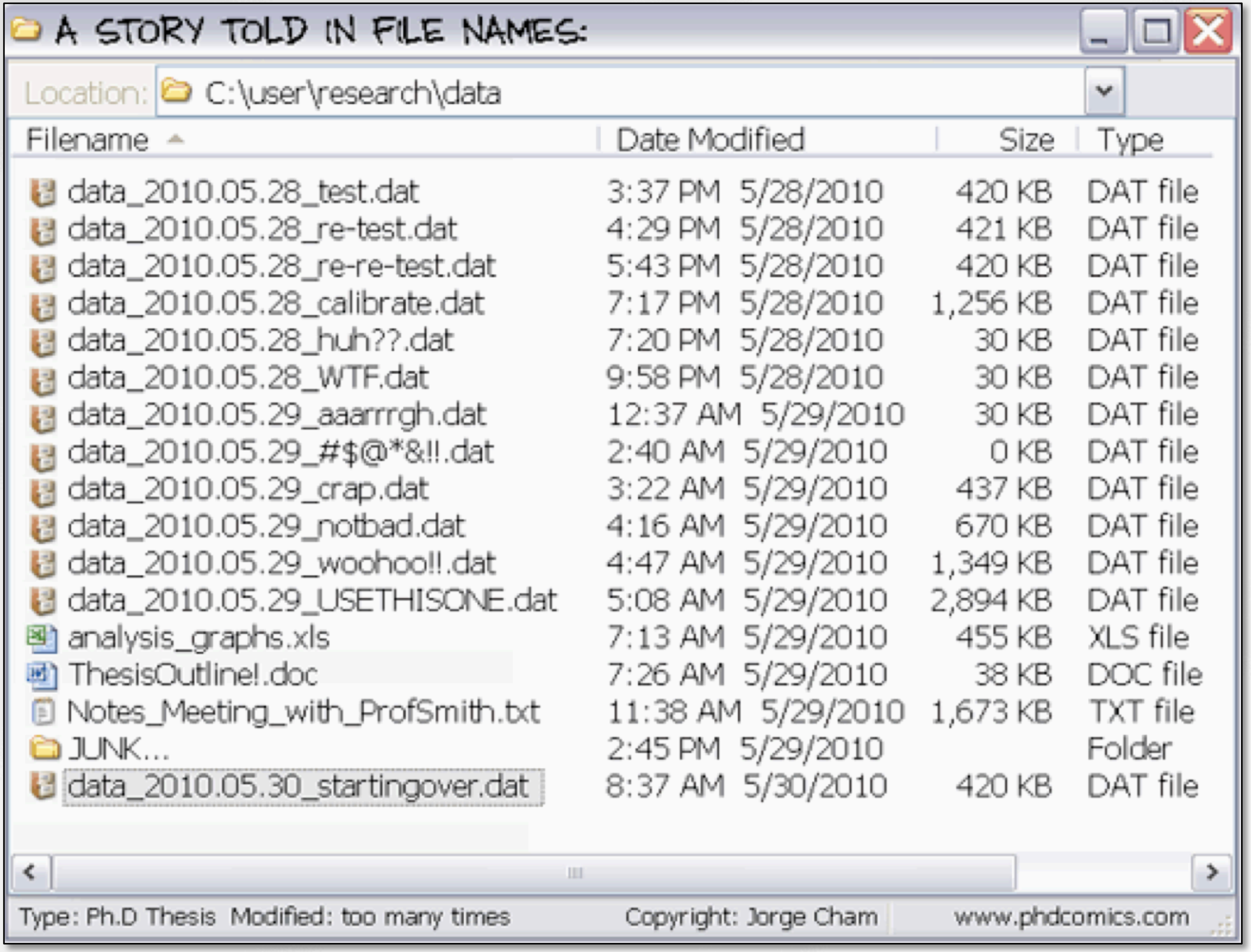
Provenance Capture in the Local Desktop



```
# CIG  Vhel  e_Vhel  r_Vhel  Dist  MType  e_MType  OptAssym  r_MType  Bmag  e_Bmag
    1  7299.0    3.0  1    96.9    5.0  1.5   1   1    14.167  0.271   0.173  0.571  0.040  13.383
    2  6983.0    6.0  2    94.7    6.0  1.5   0   1    15.722  0.324   0.255  0.278  0.031  15.157
    3                       4.0  1.5   0   1   16.057  0.507  0.246  0.354        15.457
    4  2310.0    1.0  3    31.9    3.0  1.5   0   1    12.818  0.424   0.252  0.863  0.017  11.685
    5  7865.0   10.0  3   105.9    0.0  1.5   0   1    15.602  0.364   0.225  0.131  0.118  15.128
   72  5164.0    9.0  2    68.5    5.0  1.5   1   1    14.445  0.325   0.315  0.367  0.028  13.735
```

INSTITUTO DE ASTROFÍSICA DE ANDALUCÍA, IAA-CSIC

Provenance Capture in the Local Desktop



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

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 and user-friendly

noWorkflow

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: `.noworkflow` folder
- Jupyter Notebooks Support

Capturing

- 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 inspection and analysis

- Graph based
  - Definition Provenance
  - Evolution Provenance
- Query based SQL / Prolog
  - Data Provenance
  - Execution Provenance
- Diff based
  - Evolution Provenance
  - Forward / Rewind

INSTITUTO DE ASTROFÍSICA DE ANDALUCÍA, IAA-CSIC

DEMO

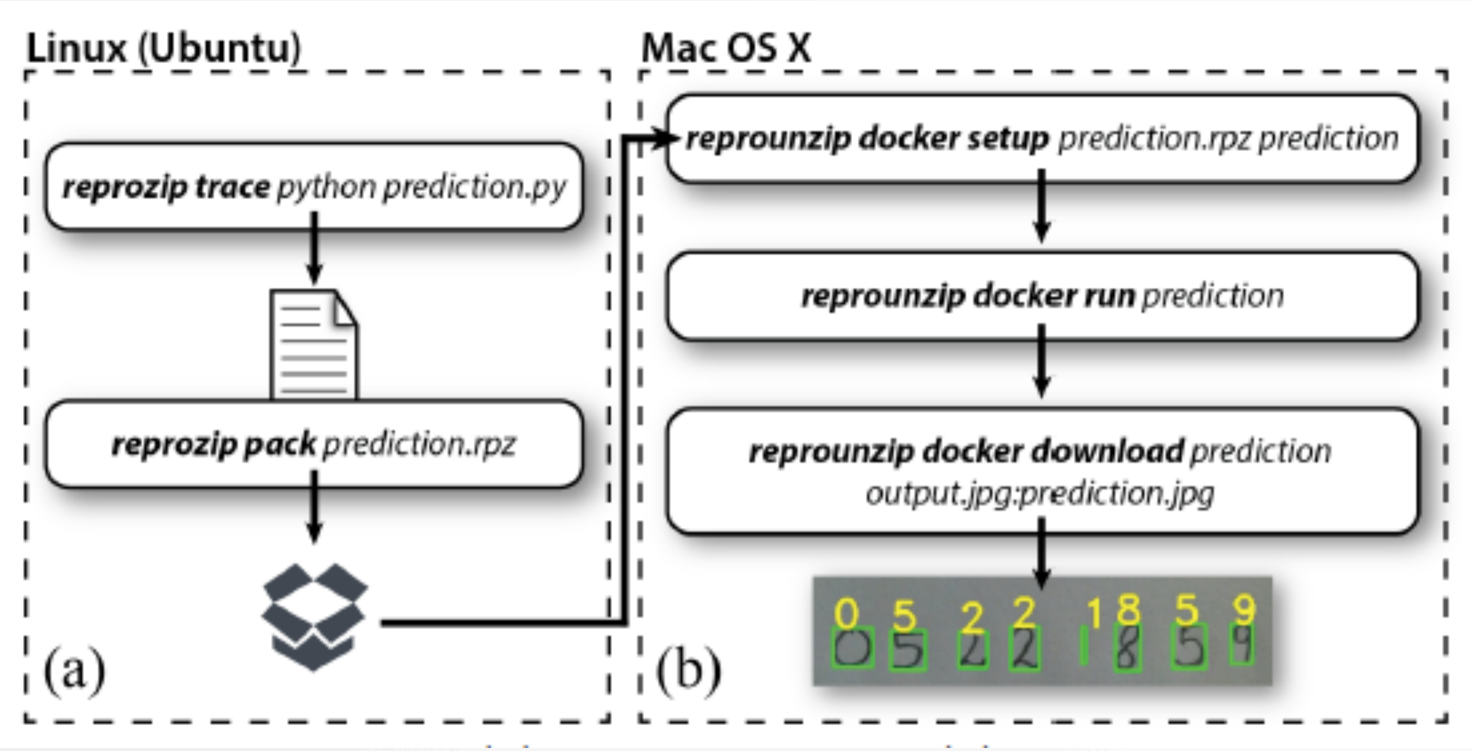
Identified use cases with gammapy

- Time profiling
- Data / files origin
- Environment inspection / switching
- Reproducibility across users
- Scripts and notebooks cells support
- Similar runs comparison…

reproZip

Creates a self-contained **package** that may be extracted and **executed across all platforms**: a reproducible experiment providing exactly the **same results/fails obtained by the packer user**.

Used in DL0 testbed simulations with HESSIO libs performed at MPIK



```
reprozip trace python prediction.py
reprozip pack prediction.rpz
reprounzip docker setup prediction.rpz prediction
reprounzip docker run prediction
reprounzip docker download prediction output.jpg:prediction.jpg
```