Using the Provenance from Astronomical Workflows

Michael Johnson, Luc Moreau, Adriane Chapman, Poshak Gandhi and Carlos Sáenz-Adán

UNIVERSITY OF Southampton

EPSRC
Engineering and Physical Sciences Research Council

Provenance in Astronomy

Most provenance is metadata contained in FITS image headers

Provenance systems for major telescopes are built for purpose usually without following any standard and the recorded provenance is not released to the astronomer

Provenance can offer substantial reductions in required storage and potentially, improvements in processing efficiency

Important due to the increasing tendency for astronomical data to be created in large scale astronomical surveys

Astronomical Surveys

The Two Micron All-Sky Survey

2MASS *1997-2001* ~ 10 TB total data set

Sloan Digital Sky Survey

SDSS *2000-2020* ~ 150 TB total data as of 2017

The Large Synoptic Survey Telescope

LSST (expected) *2022-2032* ~ 200 PB total data set

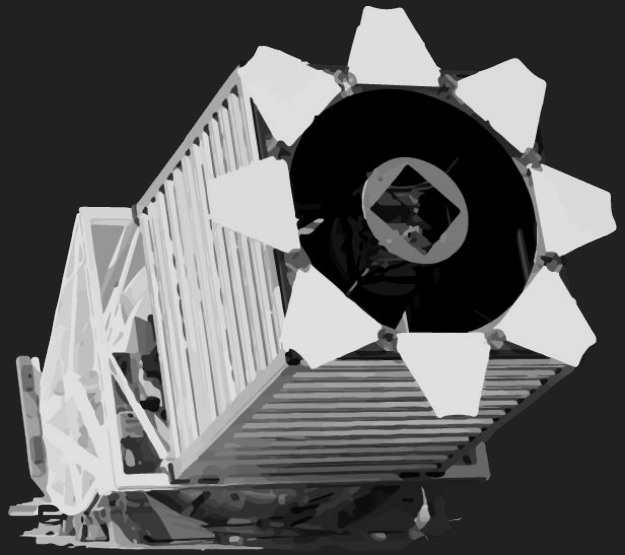


Image Credit: www.sdss.org

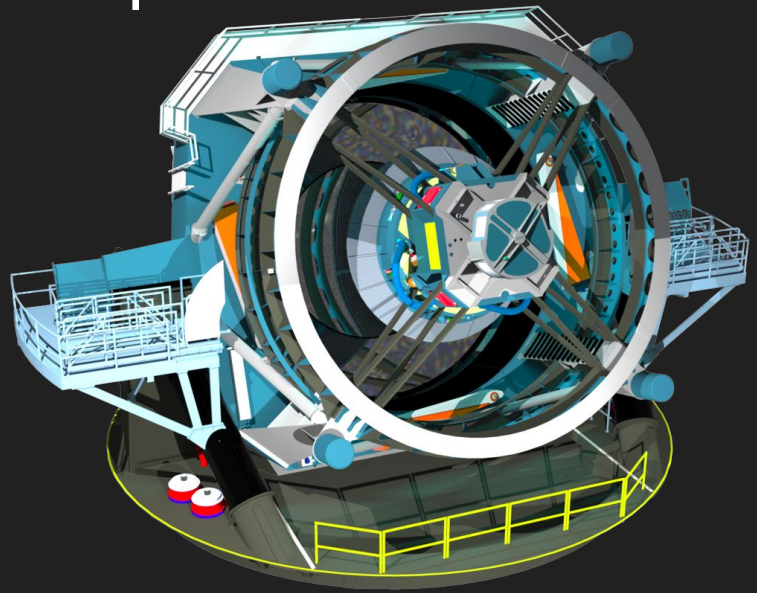
The Large Synoptic Survey Telescope

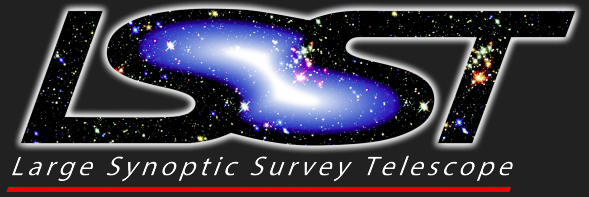
Adopting the provenance from hyper supreme cam

Will record pipeline provenance e.g. what versions of which pipelines acted on the data and on which nodes

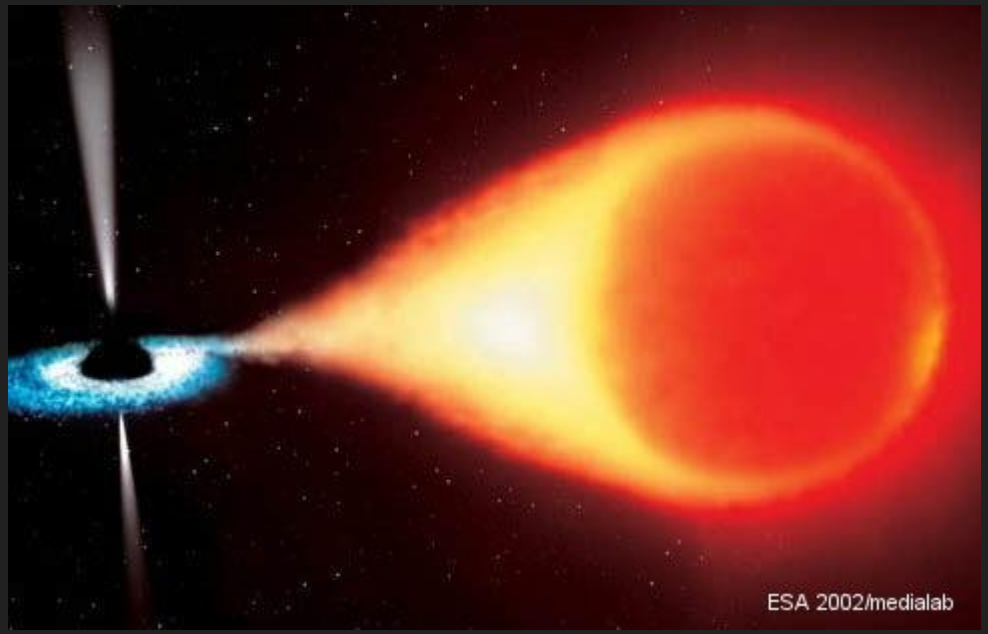
Main motivations

1) “Forensic analysis”
2) Reduction in required storage





X-ray Binaries



Low mass companion star in orbit around a compact object such as a black hole or neutron star

Compact object gravitationally attracts matter from the companion star

Matter forms an accretion disc

Only ~200 systems known

Predicted to be from thousands to millions in the Milky Way

Astronomical Images

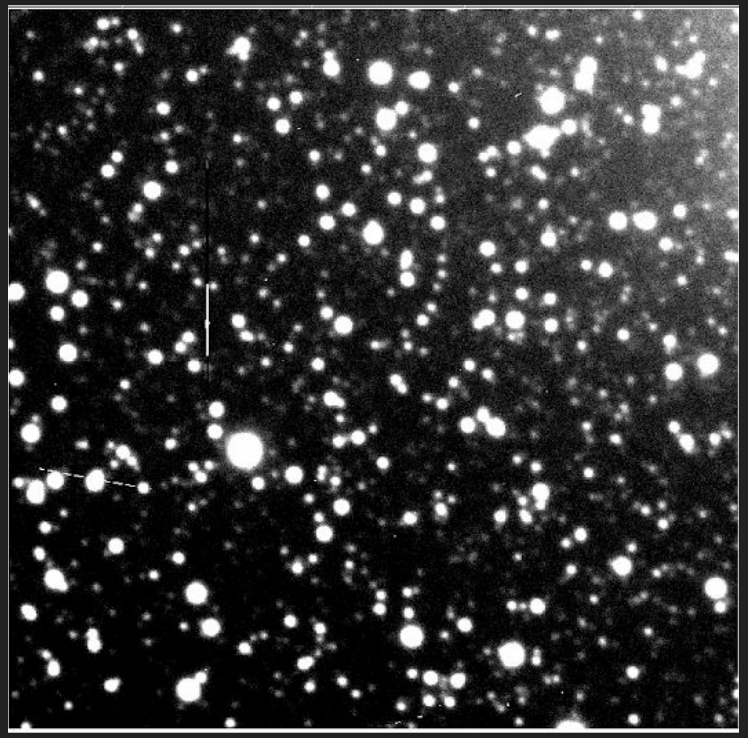


Image taken with the Faulkes Telescope

The pipeline must be able to:

-Locate the object

-Measure the brightness

-Calibrate the image (for atmospheric effects, light pollution etc.)

Differential Photometry

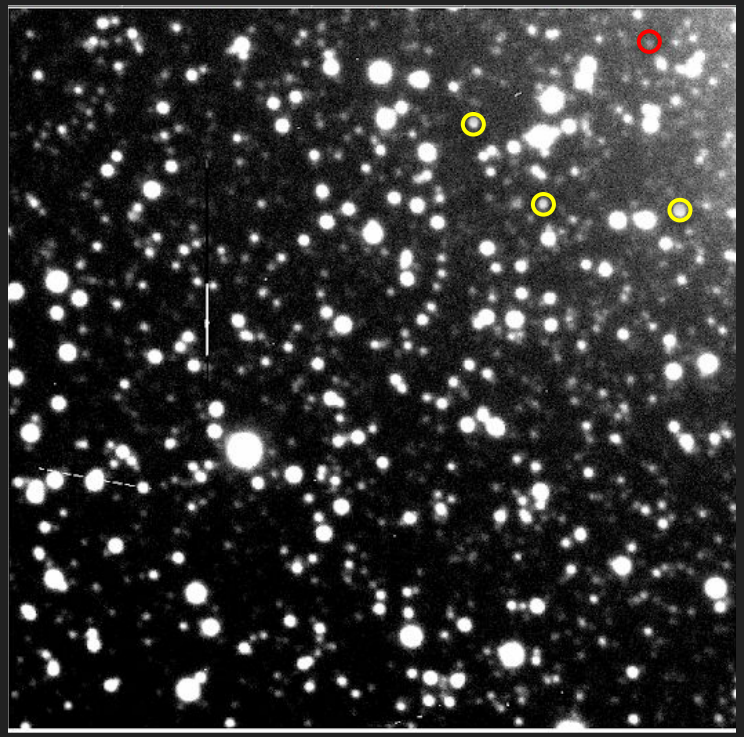


Image taken with the Faulkes Telescope

Apertures measure the brightness of the target object as well as standard objects of known brightness.

Measured brightnesses of the standard stars are then compared to their true values to create a brightness correction for the image

This correction is then used to calibrate the image for atmospheric effects and other systematics

Image Processing Pipeline

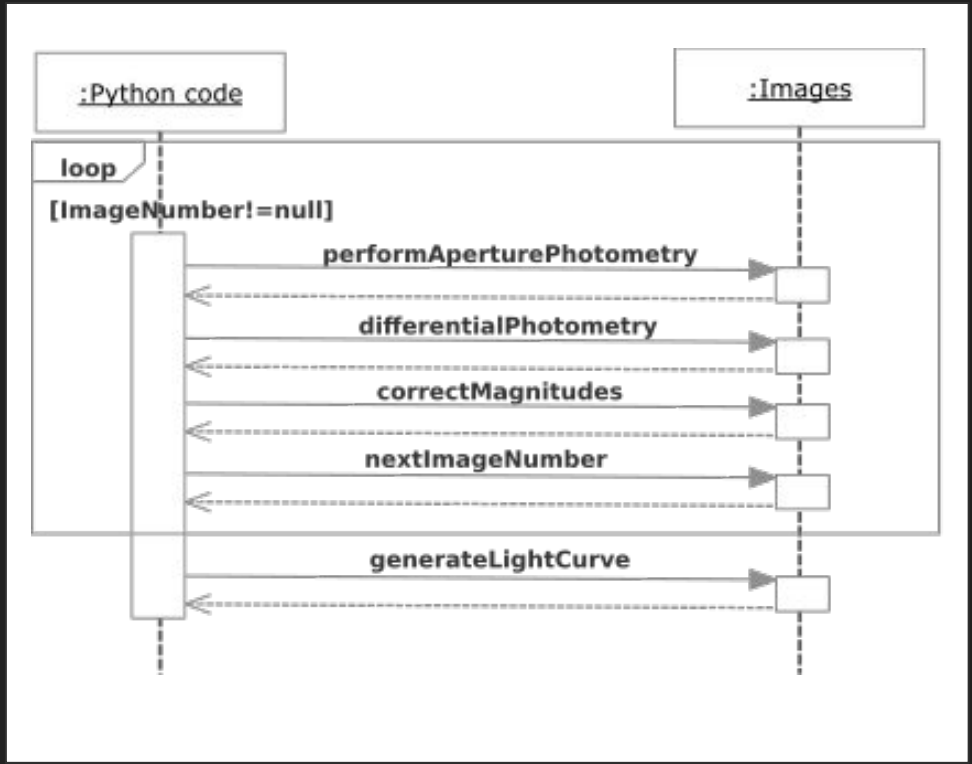
Perform aperture photometry on target and standard objects

Perform differential photometry to find the brightness correction

Correct the brightness

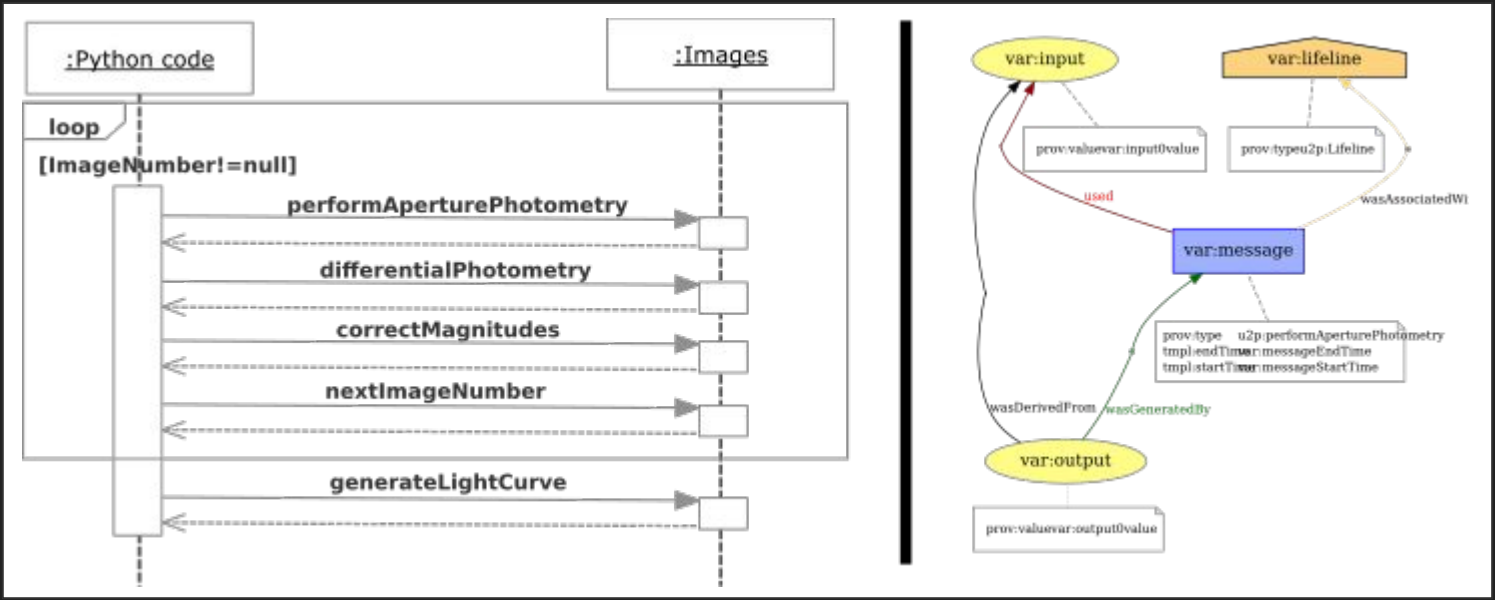
Repeat for all images

Generate a lightcurve displaying the optical variation of the LMXB



Provenance Enabling

UML2PROV (C Sáenz-Adán et al. 2018) was used to create PROV-TEMPLATES which corresponded to functions in the workflow from the workflows UML diagram

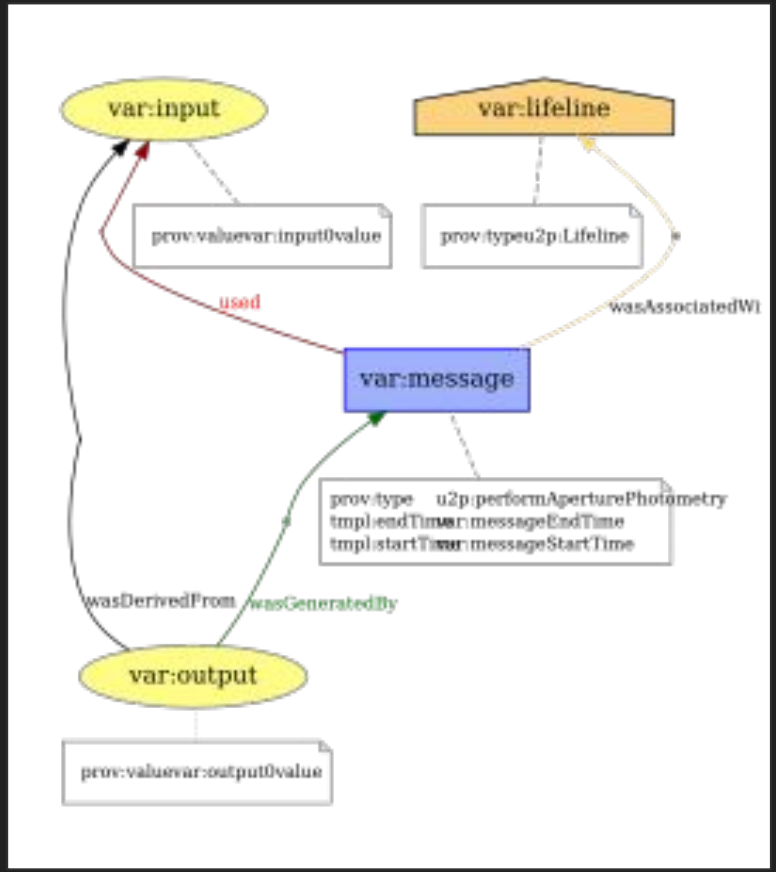


Provenance Enabling

Bindings were generated at each call of the function that contained the inputs, outputs and interactions

A python wrapper was written to automate this process

The templates were then expanded with the bindings and the product was merged to produce the full provenance



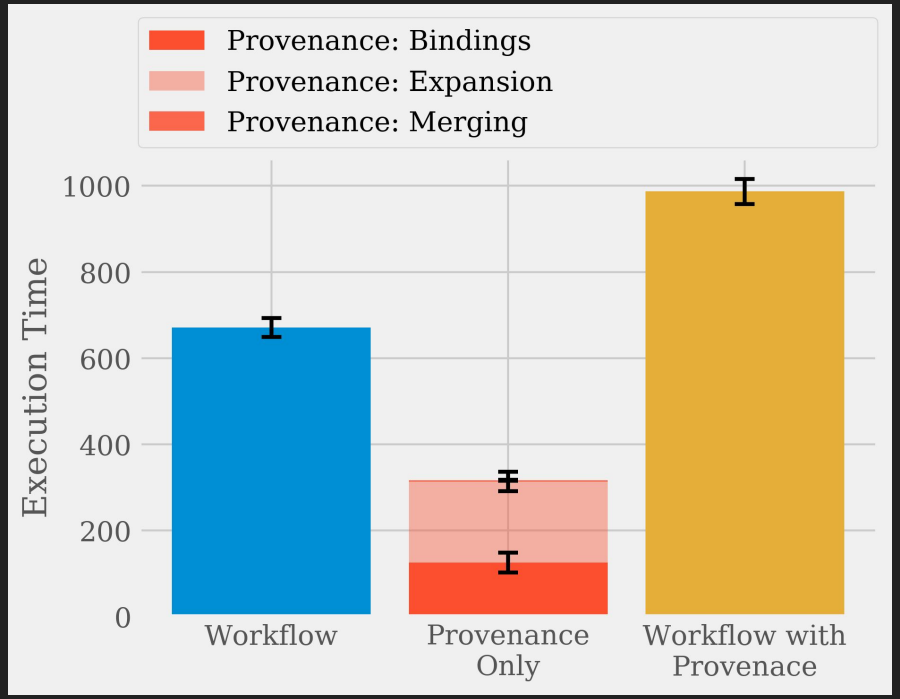
Results - Execution Evaluation

The workflow was executed 20 times with and without provenance

45% decrease in processing efficiency

Large increase in output size, however negligible when compared to the full data set

	Total Input Size	Total Output Size
Workflow Only	21 MB	20 kB
Workflow with Provenance	21 MB	546 kB

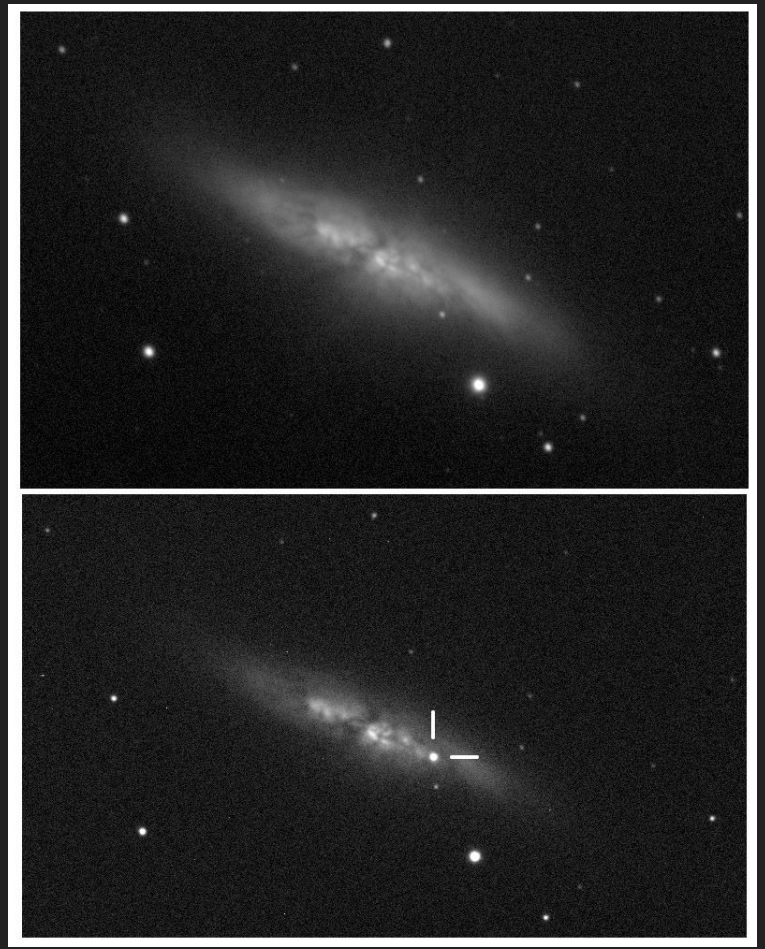


Use Case 1

USE CASE 1. *Variation Investigation* - An Astronomer, Alice, detects a change in luminosity in a star between two images taken on two different nights.

Alice *determines whether the change was a result of inconsistencies in the image processing pipeline between images.*

Speculated that this would need to be evaluated in the range of 1%, 10% or 30%

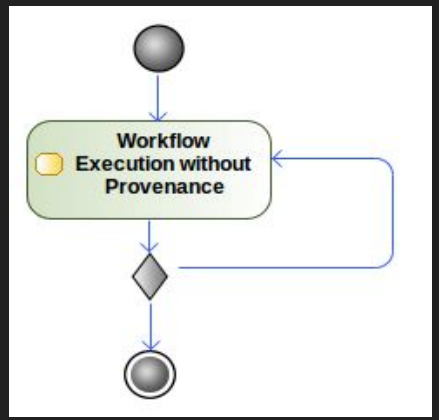


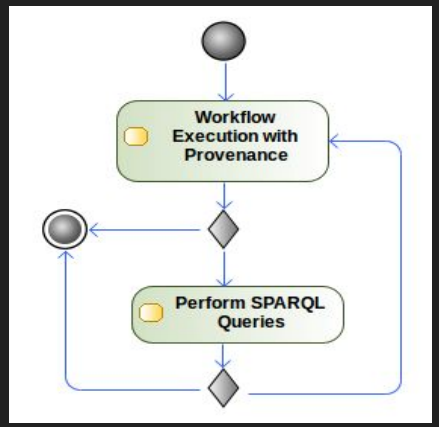
By UCL/University of London Observatory

Use Case 1 - Analysis

Without Provenance - Complete workflow re-execution, ensuring that the photometry settings are consistent throughout the series of images

With Provenance - SPARQL queries over the provenance to determine whether workflow versions and photometry settings, evaluating whether re-evaluation is necessary, if so then workflow re-execution





Use Case 2

USE CASE 2. *Calibration Propagation* - A star contained in the image, previously thought to be of standard and constant magnitude, was found to be measured incorrectly.

Alice *determines which objects used this star for calibration and recalculates the photometry for them.*

Speculated to need to be evaluated 1% of the time

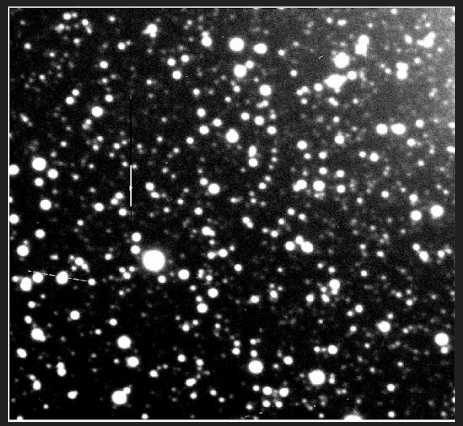


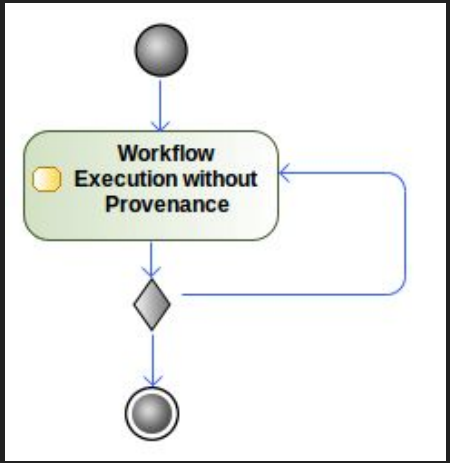
Image taken with the Faulkes Telescope

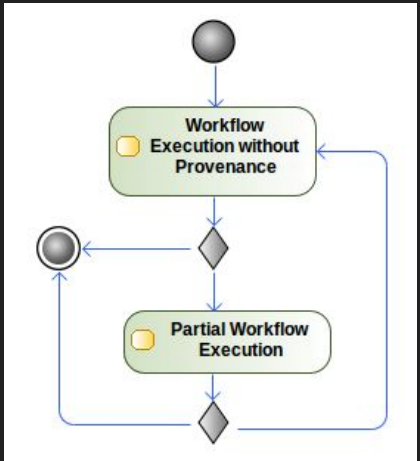
Image credit: Isadora Tatiana Nun, Harvard

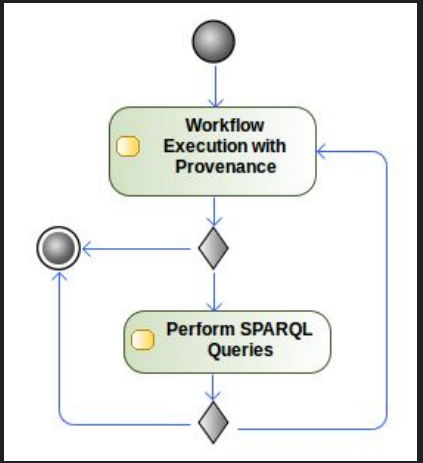
Use Case 2 - Analysis

Without Provenance - complete re-execution *or* partial re-execution to determine standard stars used, followed by re-execution if necessary

With Provenance - SPARQL queries to determine standard stars, followed by re-execution if necessary







Results - Use Case 1 Evaluation

Without provenance - the pipeline versions and photometry settings were manually set to be consistent throughout the image series and the pipeline was re-run

With provenance - SPARQL queries determined that the pipeline versions and photometry settings were consistent throughout the image series, therefore no re-running was necessary

Use Case 1 - ~99% increase in processing efficiency when evaluating with provenance

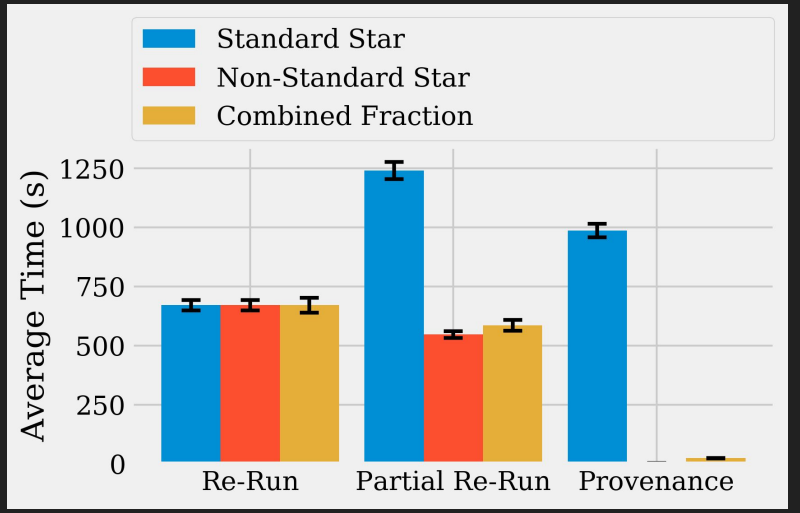
	Use Case 1 Analysis Computation Time (s)	Standard Deviation (s)	Approximate Lines of Code
Workflow Only	671	22	500
Workflow with Provenance	<1	0	10

Results - Use Case 2 Evaluation

Time required for evaluating Use Case 2 depends on whether the incorrectly measured star was used in the calibration

Standard Star and Non-Standard Star columns represent the time required for use case evaluation when the target object was and was not used in the calibration, respectively

Combined fraction combines the two evaluation times with the probability that any star in the image was used as a standard star



Use Case 2 - ~95% increase in processing efficiency when evaluating with provenance

Final Results

Combining processing costs for workflow evaluation and use case evaluation with the probability that the use cases need to be evaluated reveals a net decrease in processing efficiency of 13-44%.

	Workflow Run Time (s)	Use Case 1 Run Time (s)	Use Case 2 Run Time (s)	Total Run Time (s)
Workflow Only	671	7, 67, 201	6	684, 744, 878
Workflow with Provenance	987	<1	<1	988

Summary

An image processing pipeline was designed to measure the optical variation of astronomical objects and subsequently provenance enabled

Recording provenance introduced an increase to the initial processing cost of ~45%

Evaluating Use Case 1 and Use Case 2 with provenance increased the processing efficiency by ~99% and ~95%, respectively

A net decrease in the processing efficiency of 13-44% was found

There is the possibility to increase the total processing efficiency by introducing additional use cases