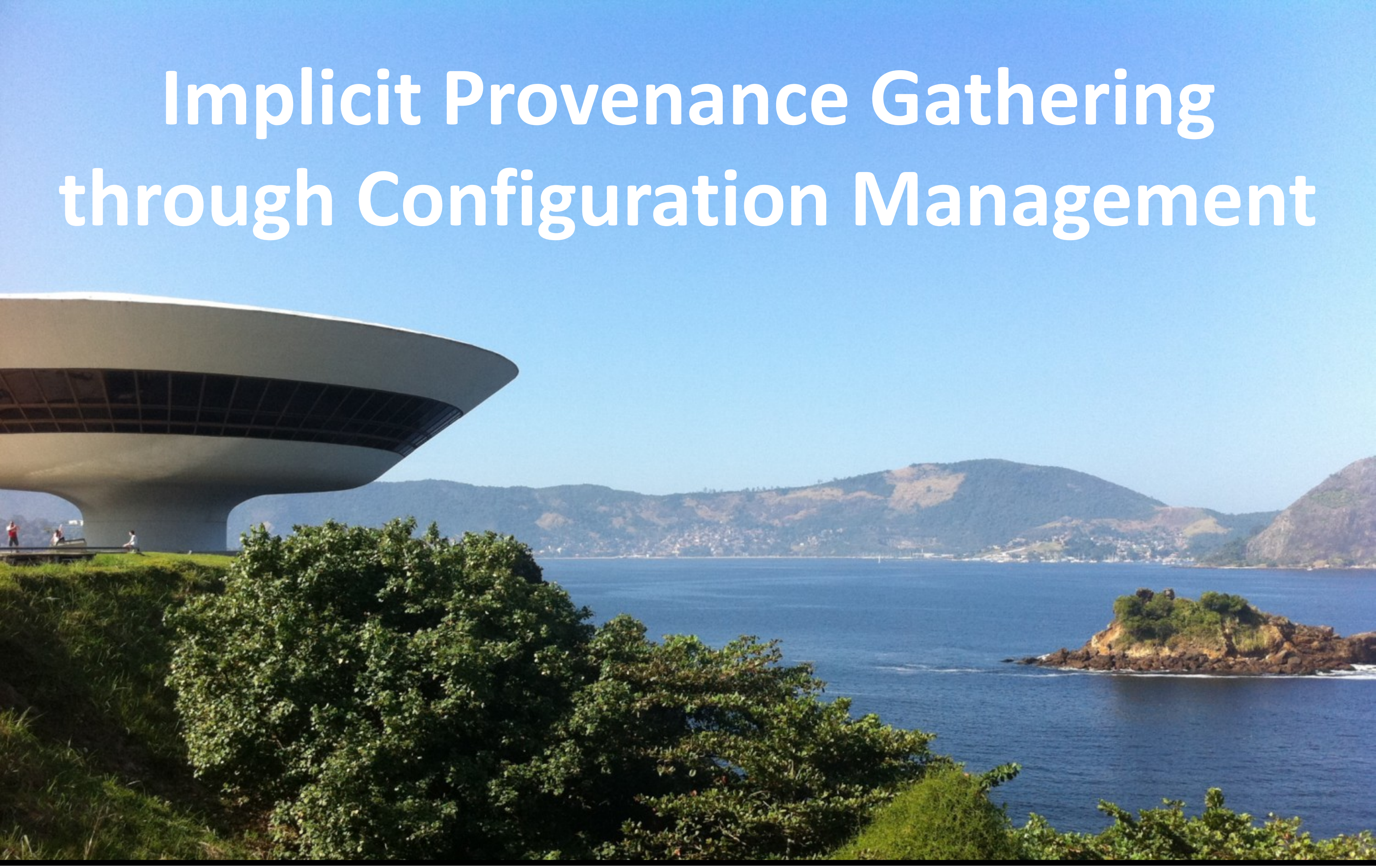
Implicit Provenance Gathering through Configuration Management



Universidade Federal Fluminense

Vitor Carvalho Neves
Vanessa Braganholo
Leonardo Murta

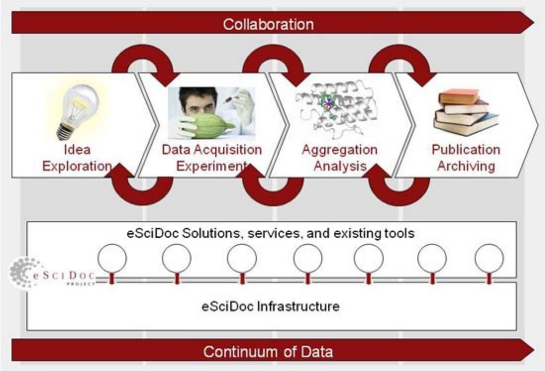
Instituto de Computação

1. Introduction and Background

Scientific experiments

System simulations

Scientific workflows



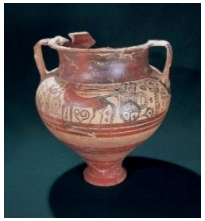
Data transformation

1. Introduction and Background

What is Data Provenance ?

1. Introduction and Background

- Provenance in arts

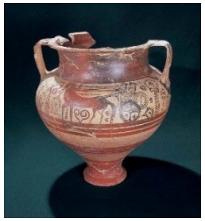


- Data provenance

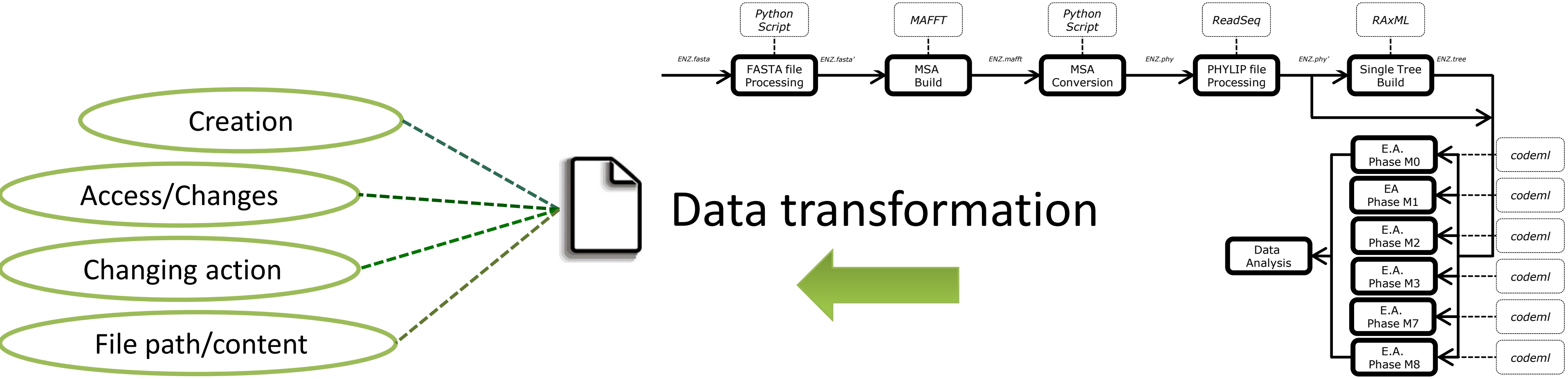
Historical information about data ownership and transformations

1. Introduction and Background

- Provenance in arts



- Provenance in the Experiment context



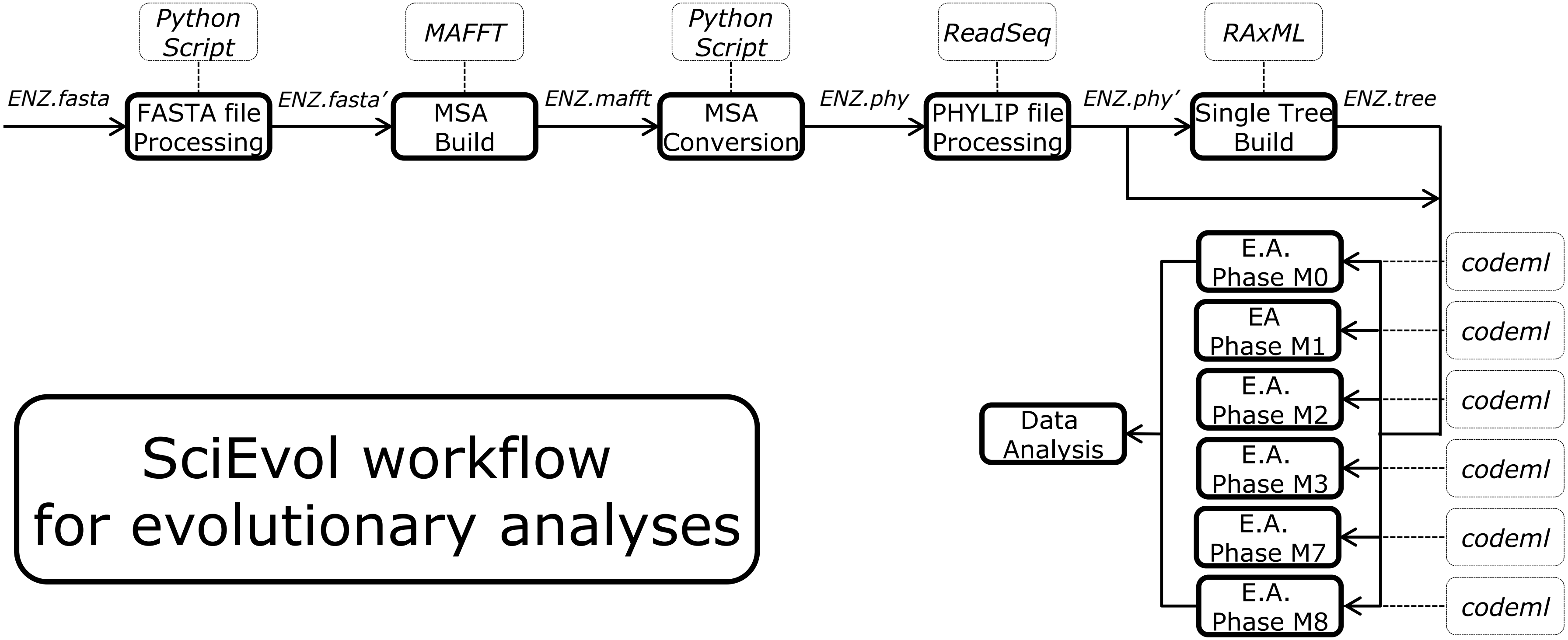
2. Problem Statement

- Provenance gathering goals:

  - Identify the action responsible for data transformation during workflow execution.

  - **Gather data files transformations, even when not explicitly specified.**

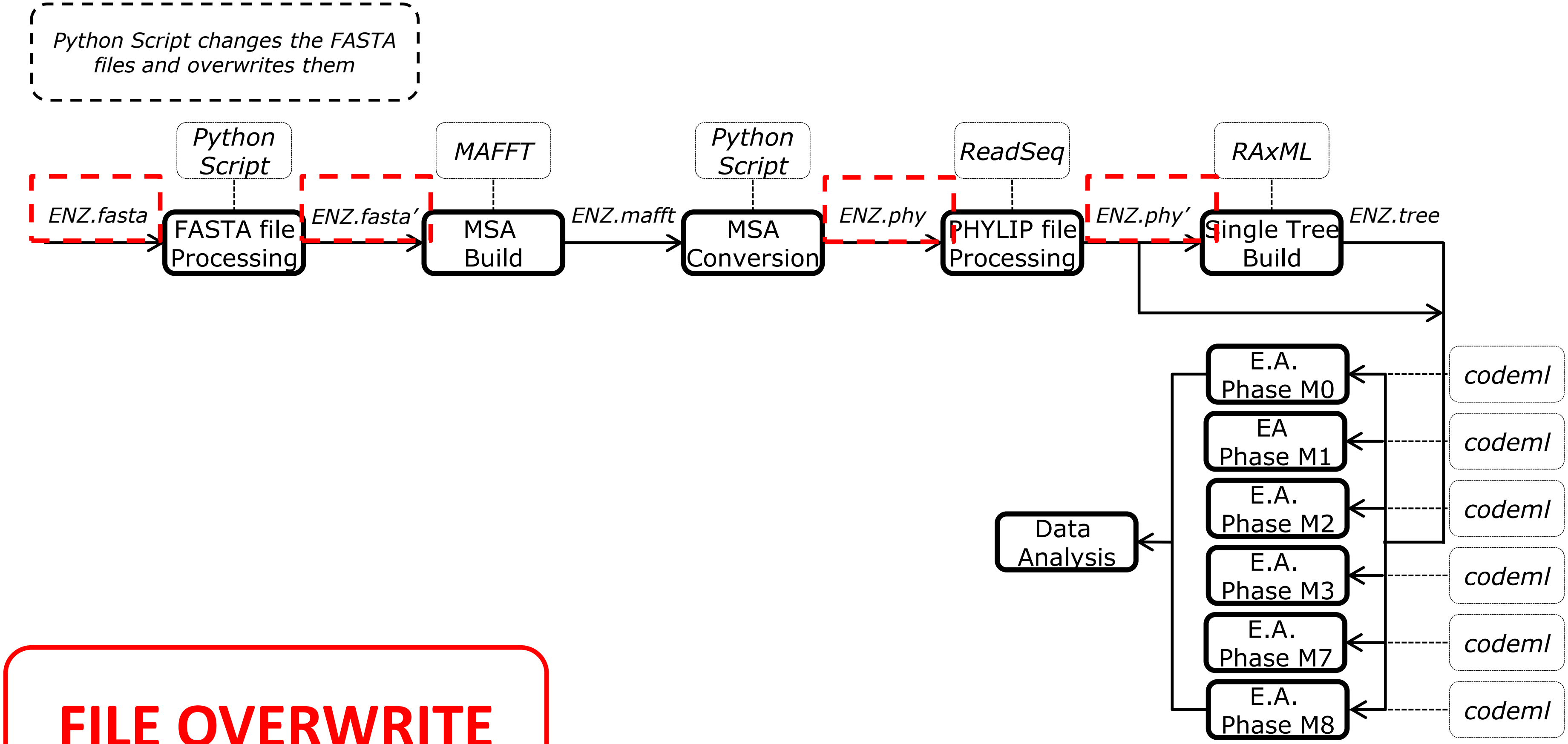
2. Problem Statement



SciEvol workflow
for evolutionary analyses

K. A. C. S. Ocaña, D. de Oliveira, F. Horta, J. Dias, E. Ogasawara, and M. Mattoso, Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific Workflow, *Advances in Bioinformatics and Computational Biology*, vol. 7409, Springer Berlin Heidelberg, 2012, p. 179-191.

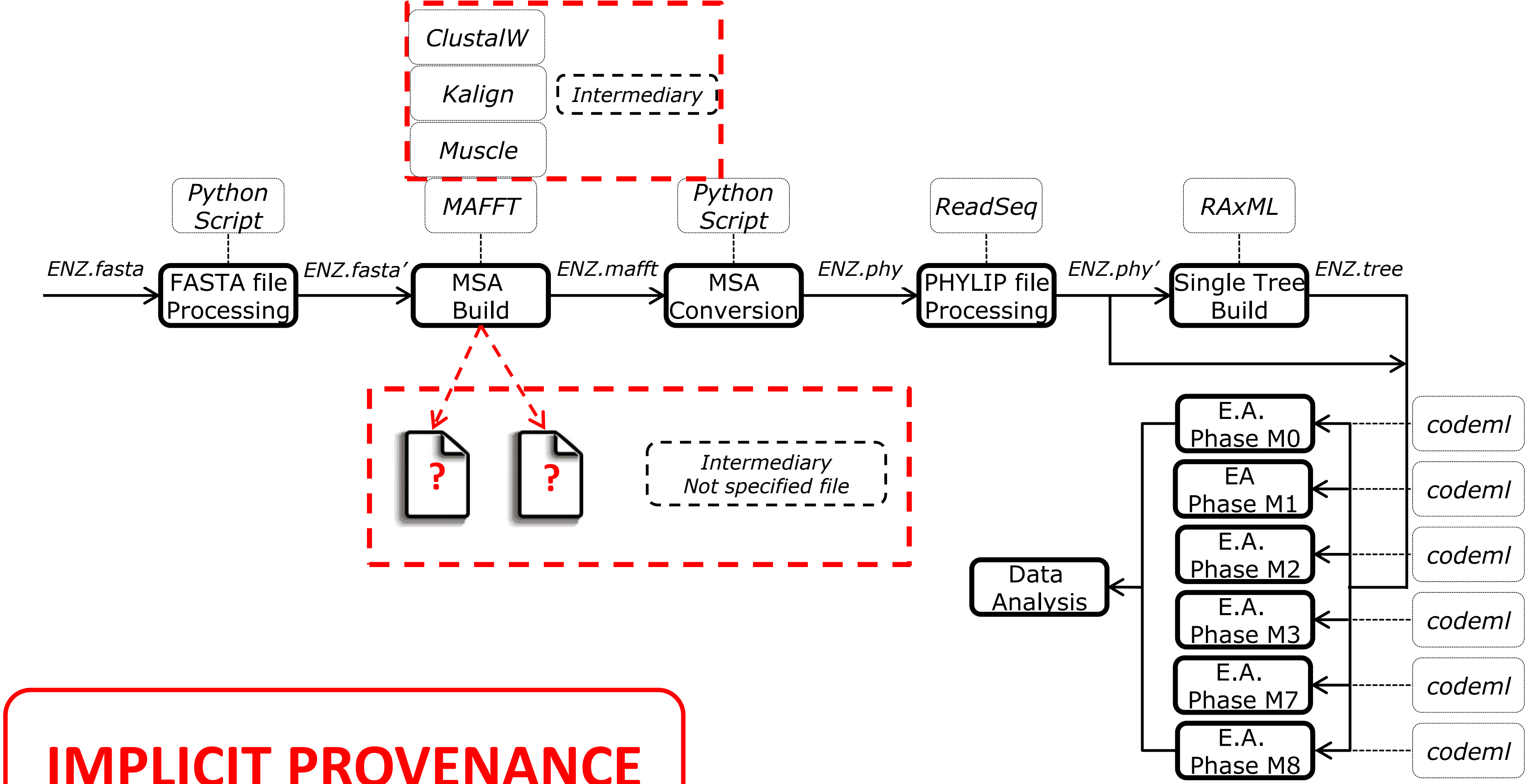
2. Problem Statement



FILE OVERWRITE

K. A. C. S. Ocaña, D. de Oliveira, F. Horta, J. Dias, E. Ogasawara, and M. Mattoso, Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific Workflow, *Advances in Bioinformatics and Computational Biology*, vol. 7409, Springer Berlin Heidelberg, 2012, p. 179-191.

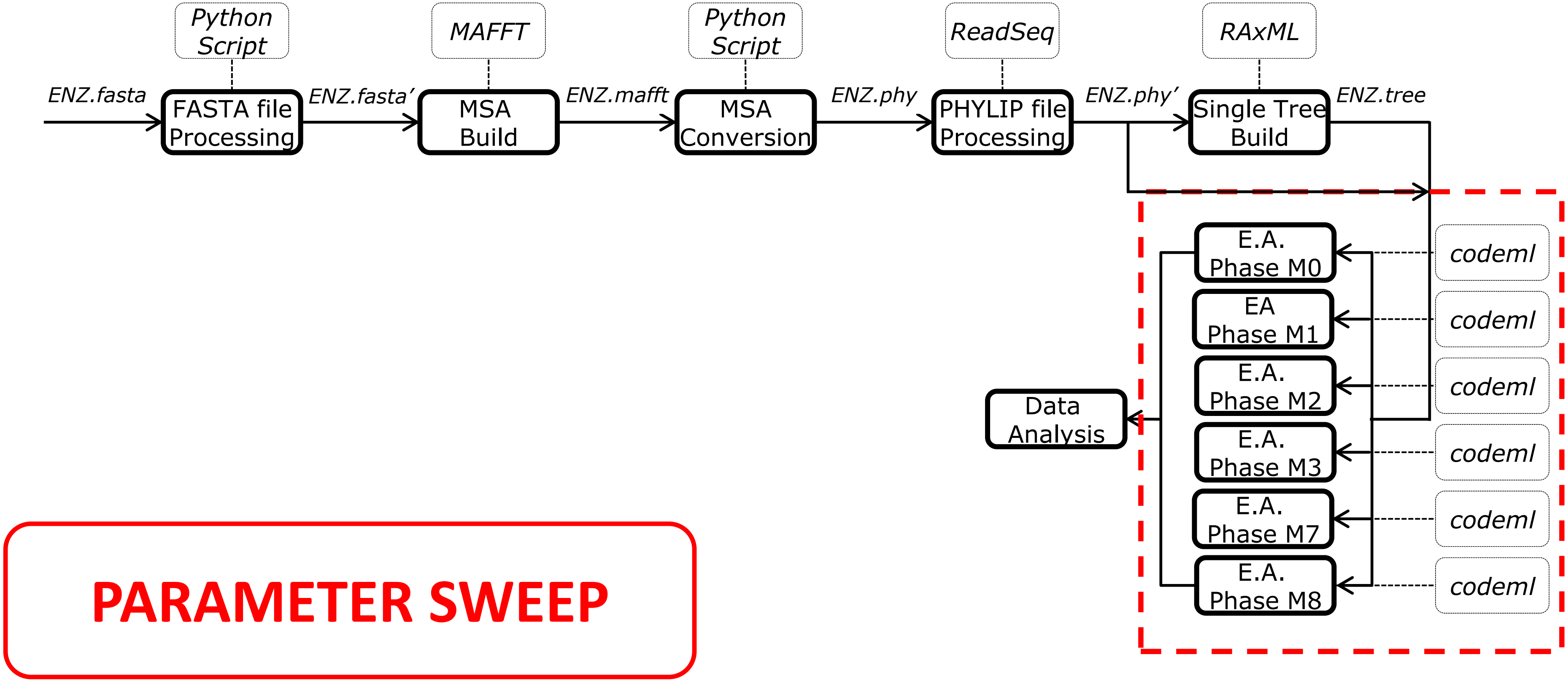
2. Problem Statement



IMPLICIT PROVENANCE

K. A. C. S. Ocaña, D. de Oliveira, F. Horta, J. Dias, E. Ogasawara, and M. Mattoso, Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific Workflow, *Advances in Bioinformatics and Computational Biology*, vol. 7409, Springer Berlin Heidelberg, 2012, p. 179-191.

2. Problem Statement



PARAMETER SWEEP

K. A. C. S. Ocaña, D. de Oliveira, F. Horta, J. Dias, E. Ogasawara, and M. Mattoso, Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific Workflow, *Advances in Bioinformatics and Computational Biology*, vol. 7409, Springer Berlin Heidelberg, 2012, p. 179-191.

3. The ProvMonitor Approach

- Provenance and CM analogy

  - □ CM perspective
    - ❖ What and Why: issue -> commit

→ workflow activity = issue to be tracked

→ experiment data = configuration item

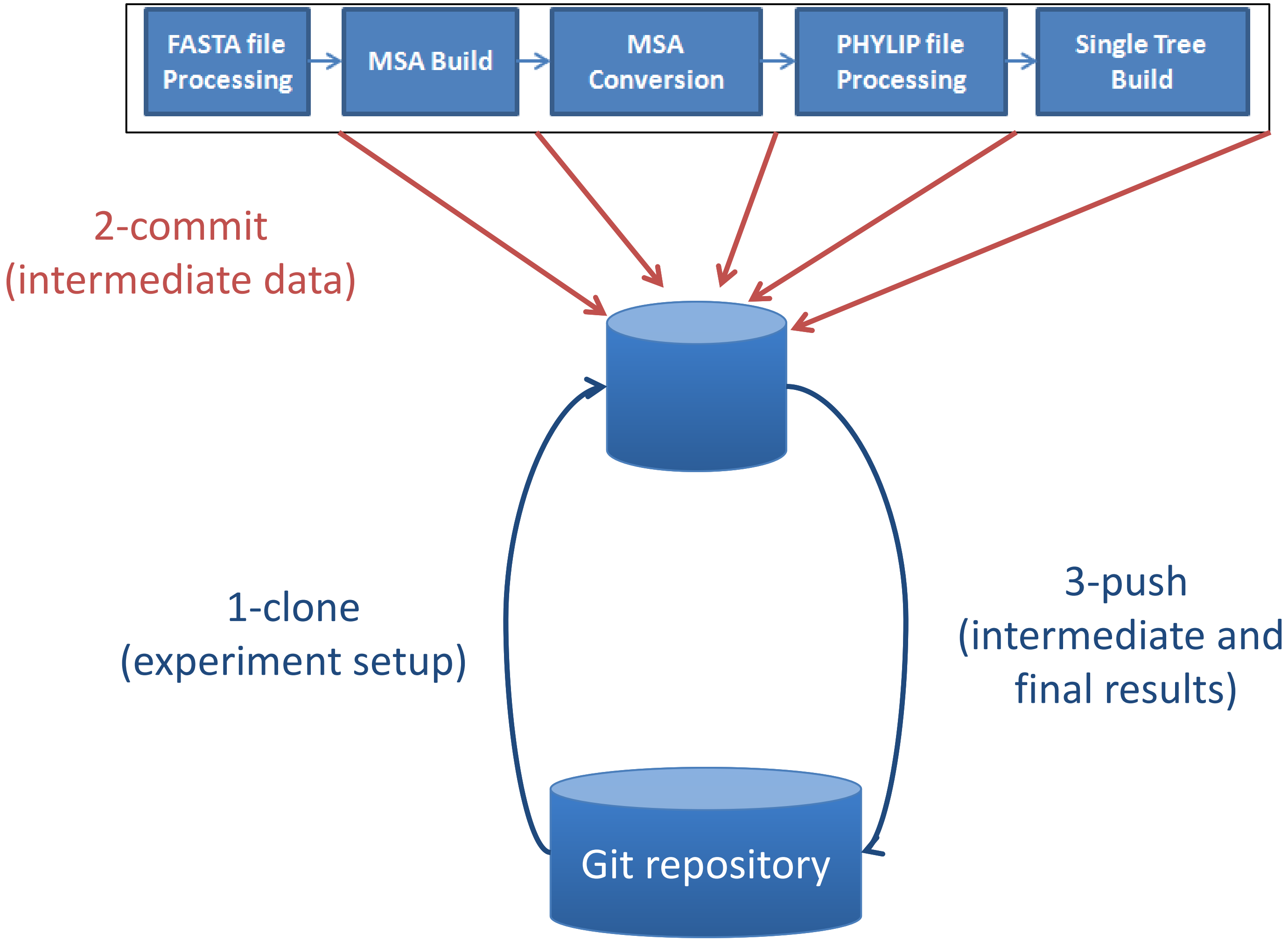
  - □ Provenance perspective
    - ❖ workflow activity -> provenance ≈ issue -> commit

3. The ProvMonitor Approach

- **Our strategy:**
  - Gather Provenance through a Configuration Management perspective.

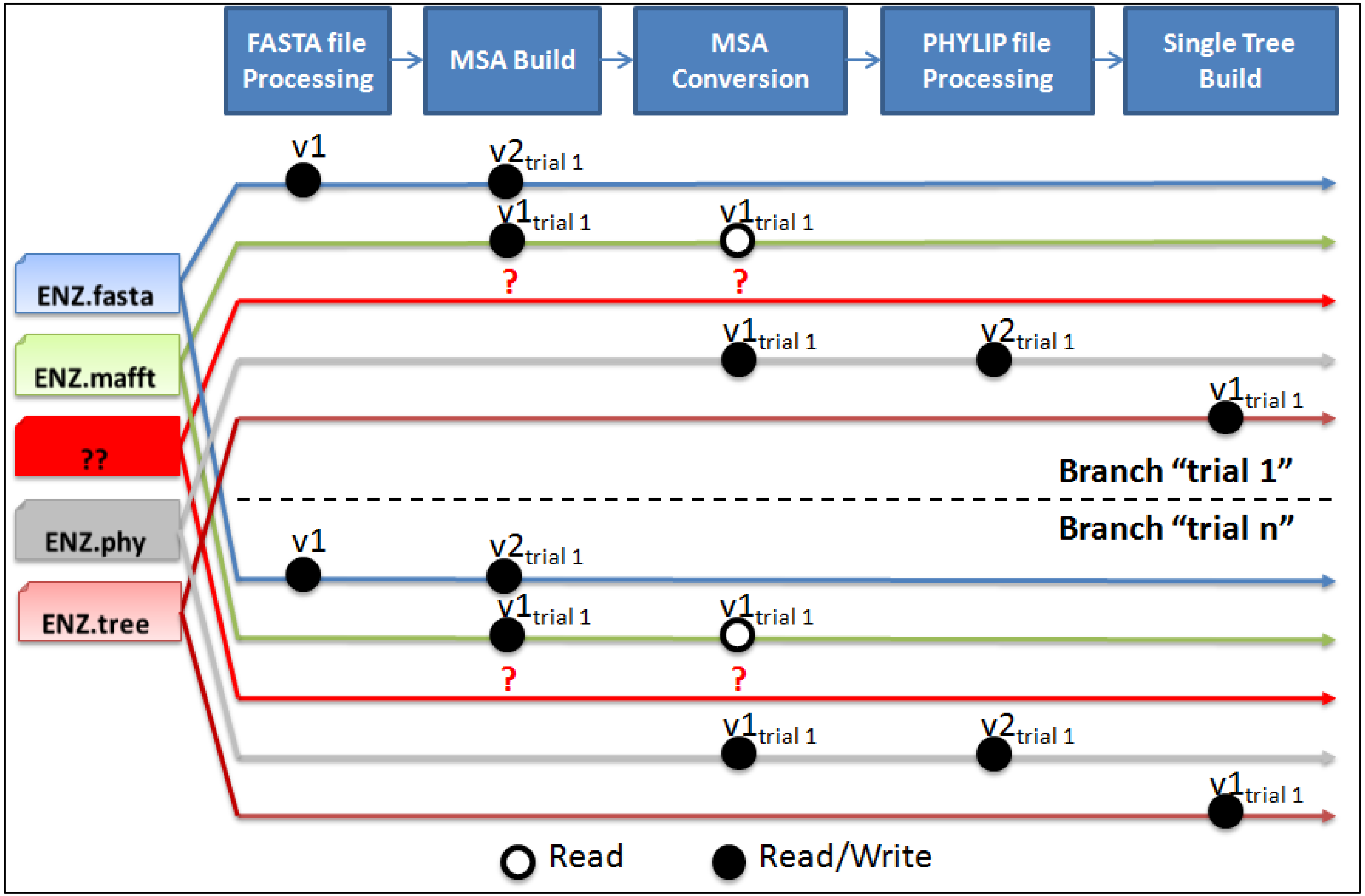
Provenance x CM	
Provenance problem	**CM feature**
File overwrite	Versioning
Implicit provenance	Workspace management
Parameter sweep	Branches

3. The ProvMonitor Approach



3. The ProvMonitor Approach

- Approach gathering and analysis perspective



4. Conclusion

- Main contributions:
  - Implicit provenance definition;
  - Provenance perspective through Configuration Management;
  - Implicit provenance gather mechanism;

4. Conclusion

- Ongoing work:
  - Experiments;
  - New analysis opportunities:
    - Inter-trial;
    - Intra-trial;

Implicit Provenance Gathering through Configuration Management

Universidade Federal Fluminense

Vitor Carvalho Neves (vcneves@ic.uff.br)
Vanessa Braganholo
Leonardo Murta

Instituto de Computação