TaPP '11

3rd USENIX Workshop on the Theory and Practice of Provenance

June 20–21, 2011 | Heraklion, Crete, Greece

usenix

Sponsored by USENIX in cooperation with ACM SIGMOD, ACM SIGPLAN, and the W3C Greece Office

Provenance Exchange, Integration and Querying

Marta Mattoso

Federal University of Rio de Janeiro, Brazil

UFRJ COPPE

Provenance
Exchange, Integration and Querying

Contributors:

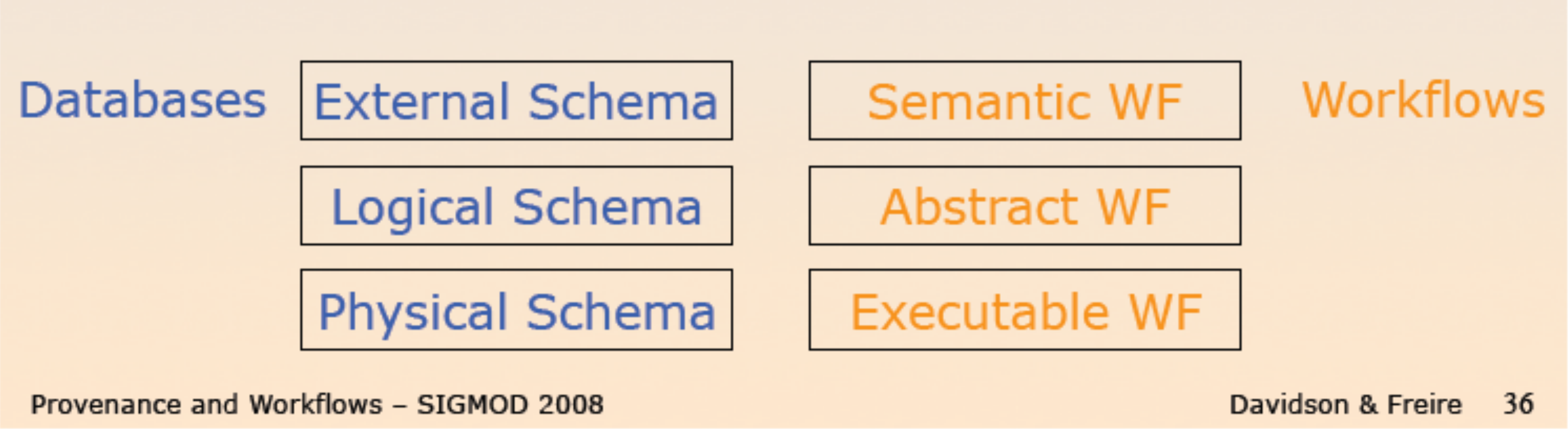
- M. David Allen, Adriane Chapman, Barbara Blaustein, Len Seligman
  [5 Getting It Together: Enabling Multi-organization Provenance Exchange]
- Anderson Marinho, Marta Mattoso, Claudia Werner, Vanessa Braganholo and Leonardo Murta
  [33 Challenges in managing implicit and abstract provenance data: experiences with ProvManager]
- Luiz M. R. Gadelha Jr., Marta Mattoso, Michael Wilde, Ian Foster
  [26 Provenance Query Patterns for Many-Task Scientific Computing]

Importance of provenance in Science

- Interpret and reproduce data
- Understand the experiment and chain of reasoning that was used in the production of a result
- Verify that an experiment was performed according to acceptable procedures
- Identify what were the inputs to an experiment and where they came from
- Assess data quality
- Track who performed an experiment and who is responsible for its results (patents)

Provenance is as (or more!) important as the results (Davidson, Freire, Provenance and Workflows- SIGMOD 2008)

Provenance along Wf levels



Databases	External Schema	Semantic WF	Workflows
	Logical Schema	Abstract WF	
	Physical Schema	Executable WF	

Provenance and Workflows – SIGMOD 2008

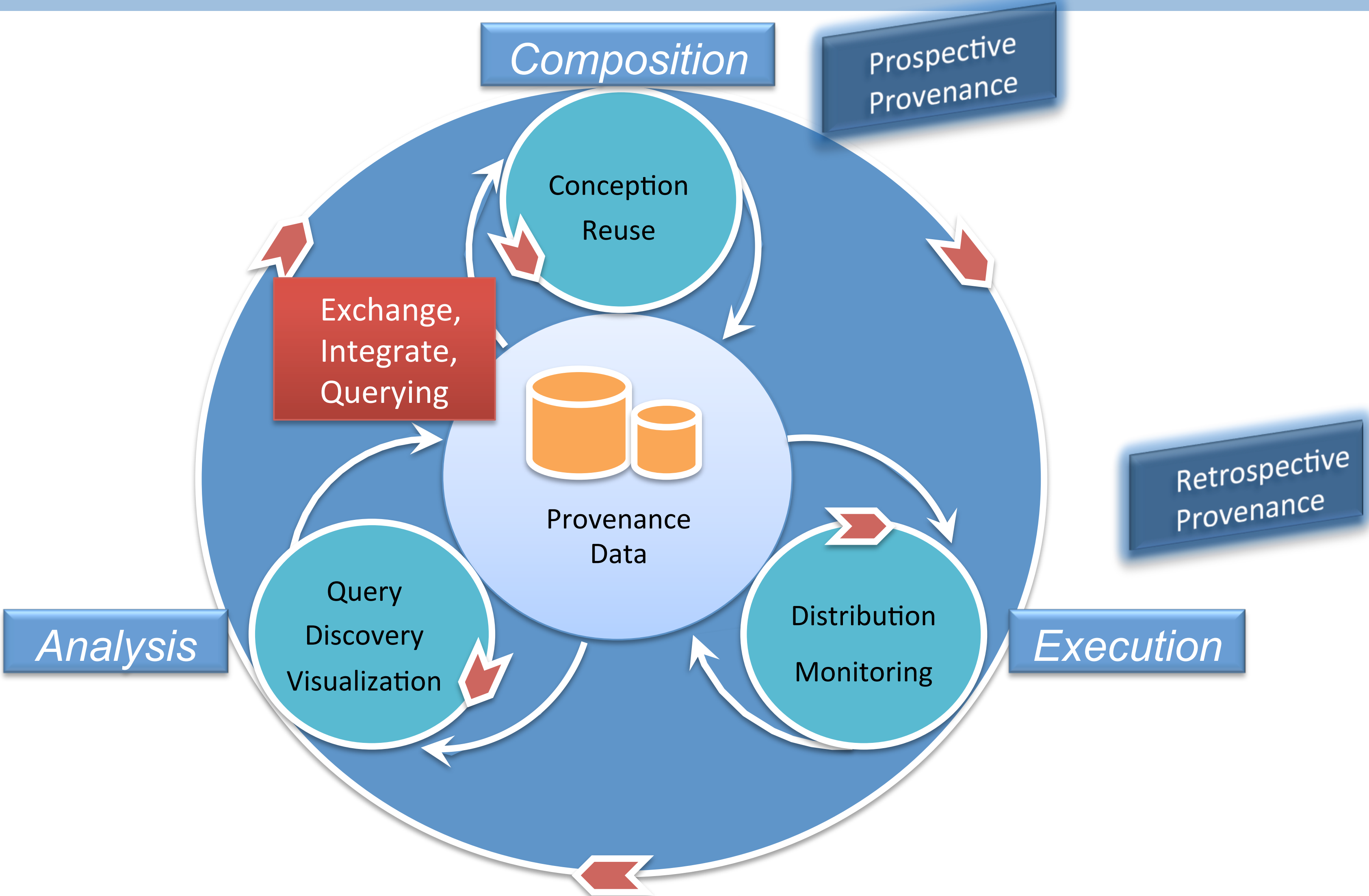
Davidson & Freire 36

Provenance can support analyzing scientific experiments

- Before execution:
  - What programs may be used? Is there any alternative methodology to explore?
  - Is there any dependency between activities? Which activities are mandatory?
- After execution:
  - What were the parameters used in the critical result ?
  - What were the scientific workflow activities used to obtain such result?
  - Where are the output files generated by the distributed activity A using the parameters P?
  - How many times the activity A in version V was used in the experiment E?

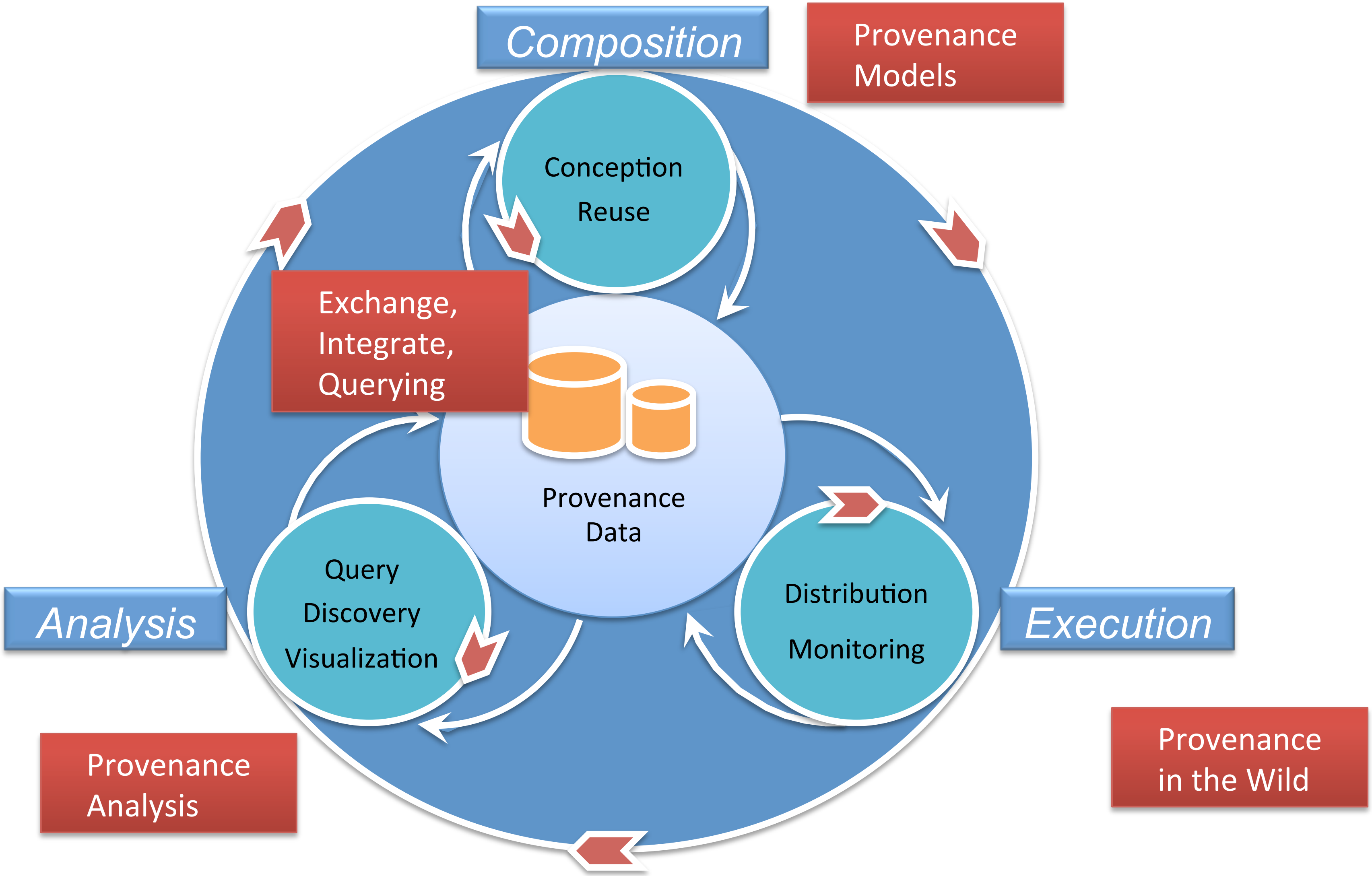
all these queries are related to the ability of reproducing and validating a scientific experiment

Experiment Life Cycle*

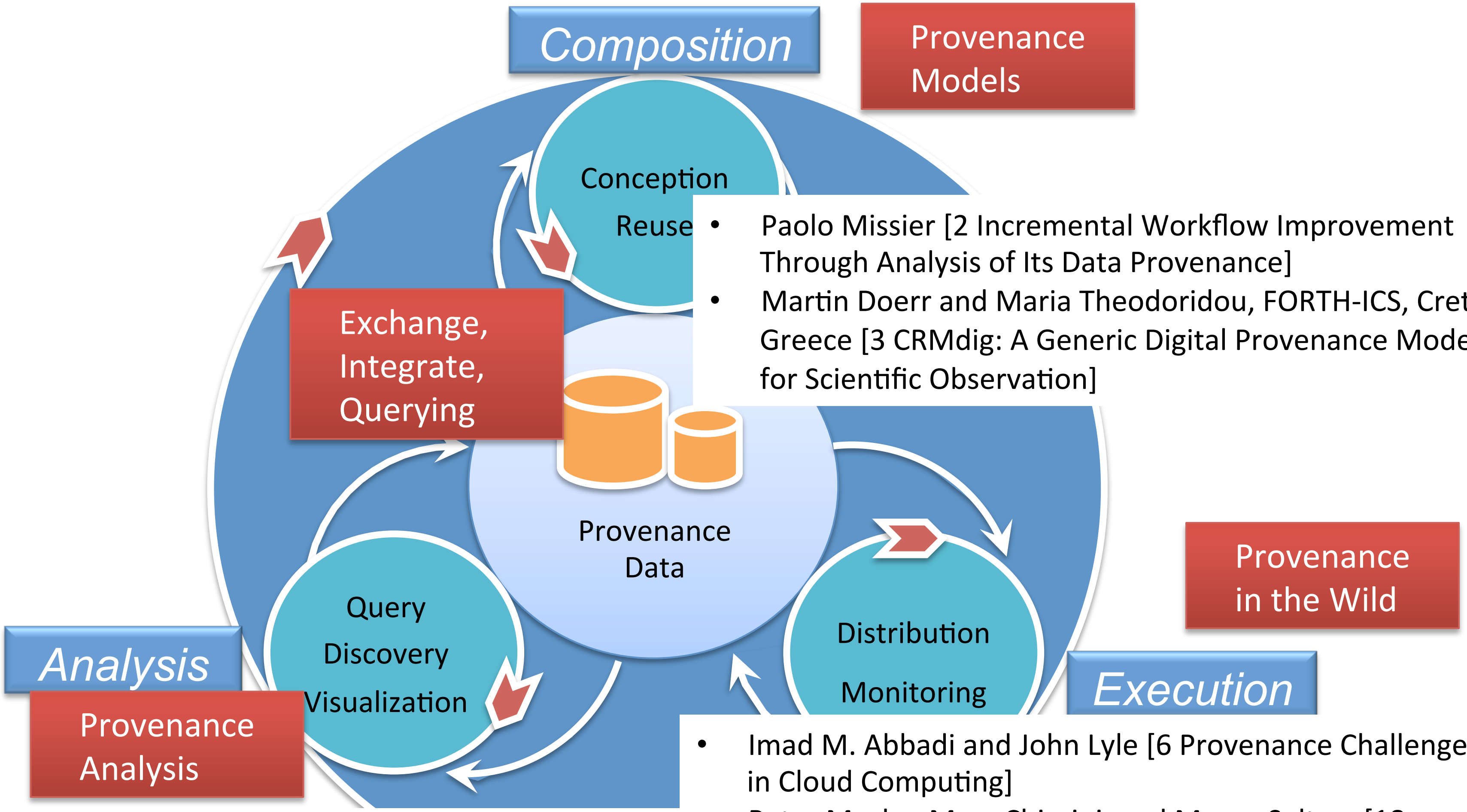


*Mattoso et al, 2010 - Towards Supporting the Life Cycle of Large Scale Scientific Experiments. IJBPIM

Experiment Life Cycle &TaPP Sessions



Experiment Life Cycle &TaPP Papers

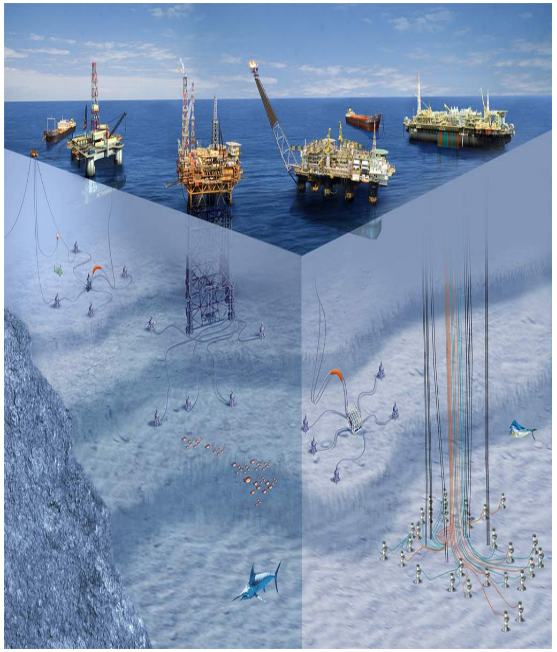


- Paolo Missier [2 Incremental Workflow Improvement Through Analysis of Its Data Provenance]
- Martin Doerr and Maria Theodoridou, FORTH-ICS, Crete, Greece [3 CRMdig: A Generic Digital Provenance Model for Scientific Observation]

- Imad M. Abbadi and John Lyle [6 Provenance Challenges in Cloud Computing]
- Peter Macko, Marc Chiarini, and Margo Seltzer [18 Collecting Provenance via the Xen Hypervisor]
- Elaine Angelino, Uri Braun, David A. Holland, Peter Macko, Daniel Margo, and Margo Seltzer [23 Provenance Integration Requires Reconciliation]

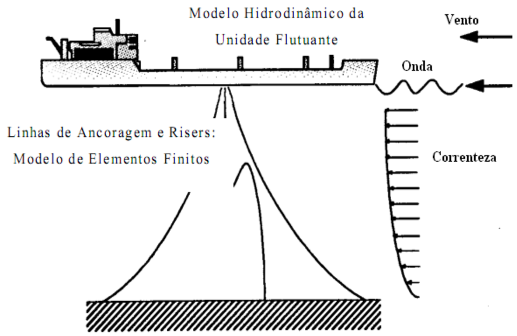
- Reng Zeng, Xudong He, Jiafei Li, Zheng Liu, W.M.P. van der Aalst [1 A Method to Build and Analyze Scientific Workflows from Provenance through Process Mining]

Risers' fatigue analysis in oil elevation from ultra-deep waters following a coupled analysis

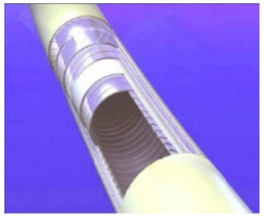


Input Data to simulate Movements:
Waves, wind, currents, batimetryDados de onda
vento, correnteza,
bathymetry, etc. :

1. Coupled movement Analysis (TPN or Prosim)

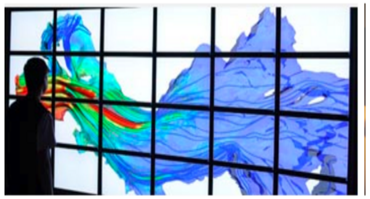


Generates large quantity of Data ...
(finite element meshes)



2. ... To do structural Analysis ofRisers (ANFLEX)

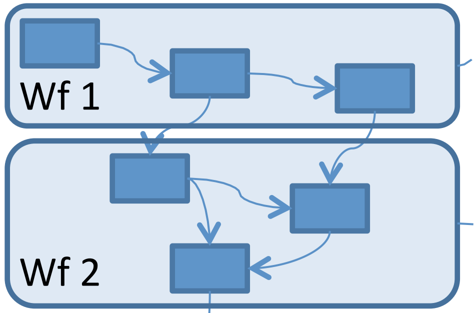
3. Results are analyzed POSFAL

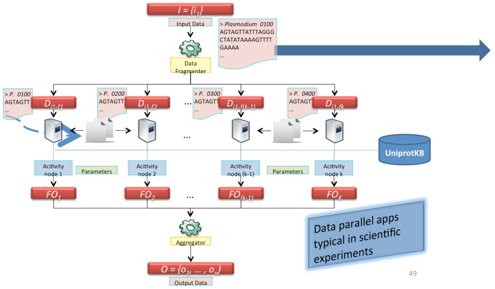


Estimate risers lifetime

Distributed Provenance

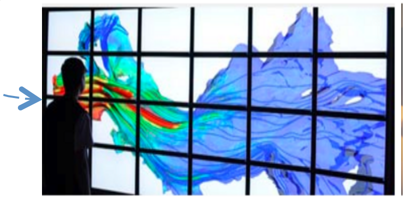
Scientific experiment

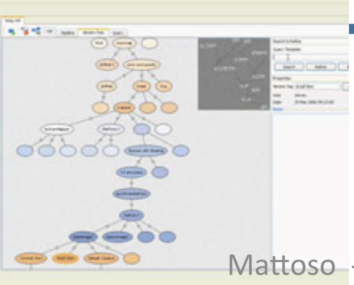




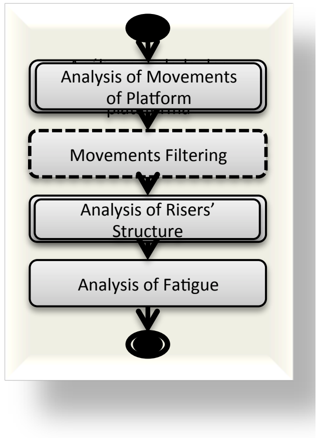
Data parallel apps typical in scientific experiments

Offline analysis (vis cave)





Provenance Systems

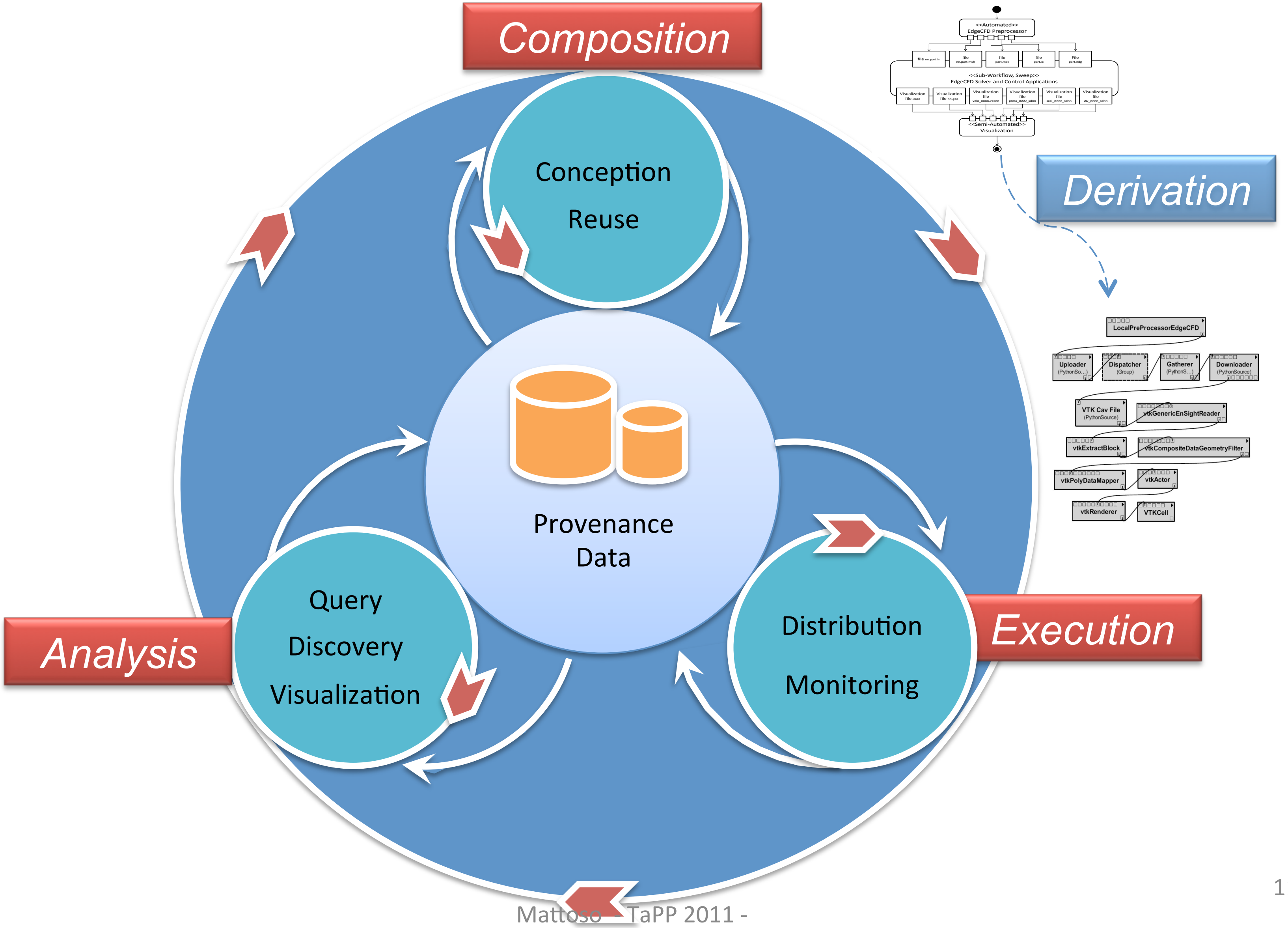

Analysis of Movements of Platform
Movements Filtering
Analysis of Risers' Structure
Analysis of Fatigue

Publish Experiment and Workflow

Sub-workflow parallel execution in HPC clusters, clouds

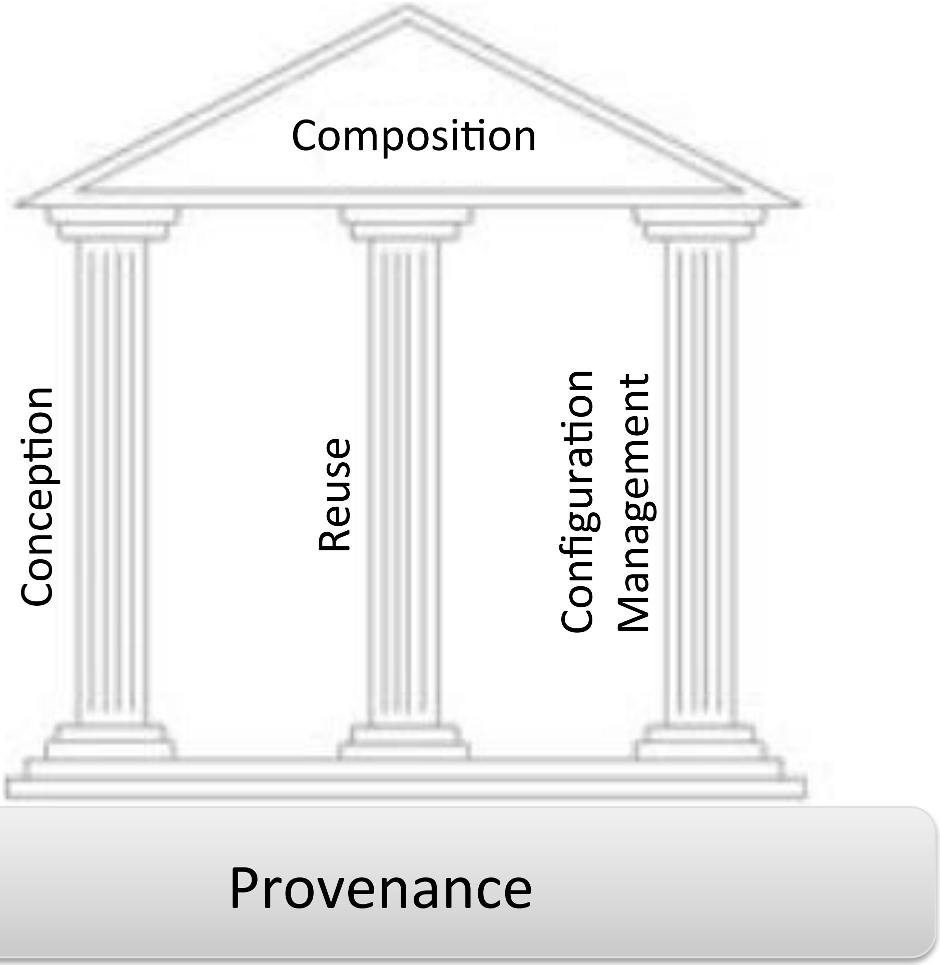
Visualize and Share provenance data with others scientists

Some aspects of Composition



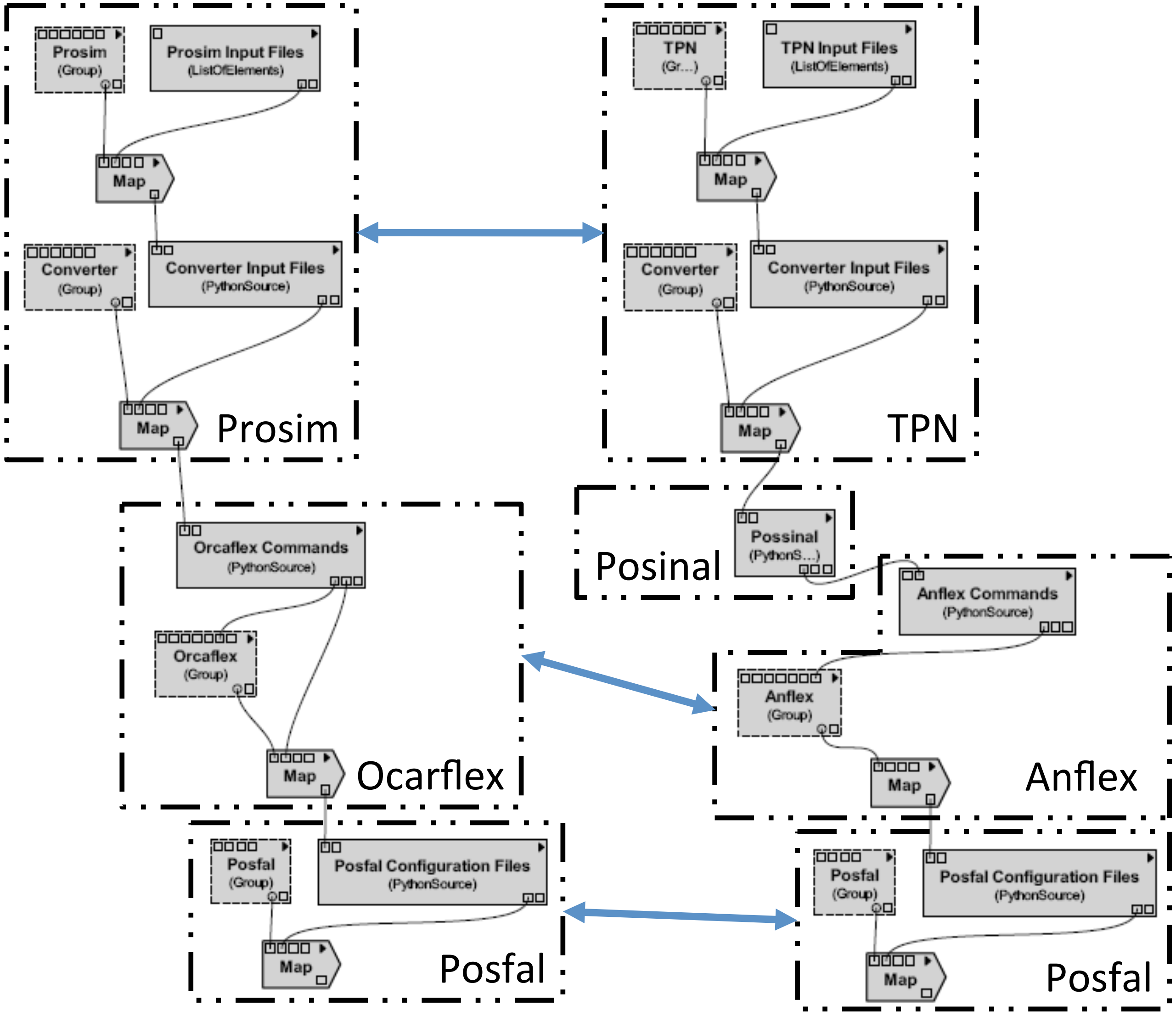
Pillars of composition

Provenance is orthogonal to those pillars and it is generated in each one of them



The composition should be supported in scientific experiments

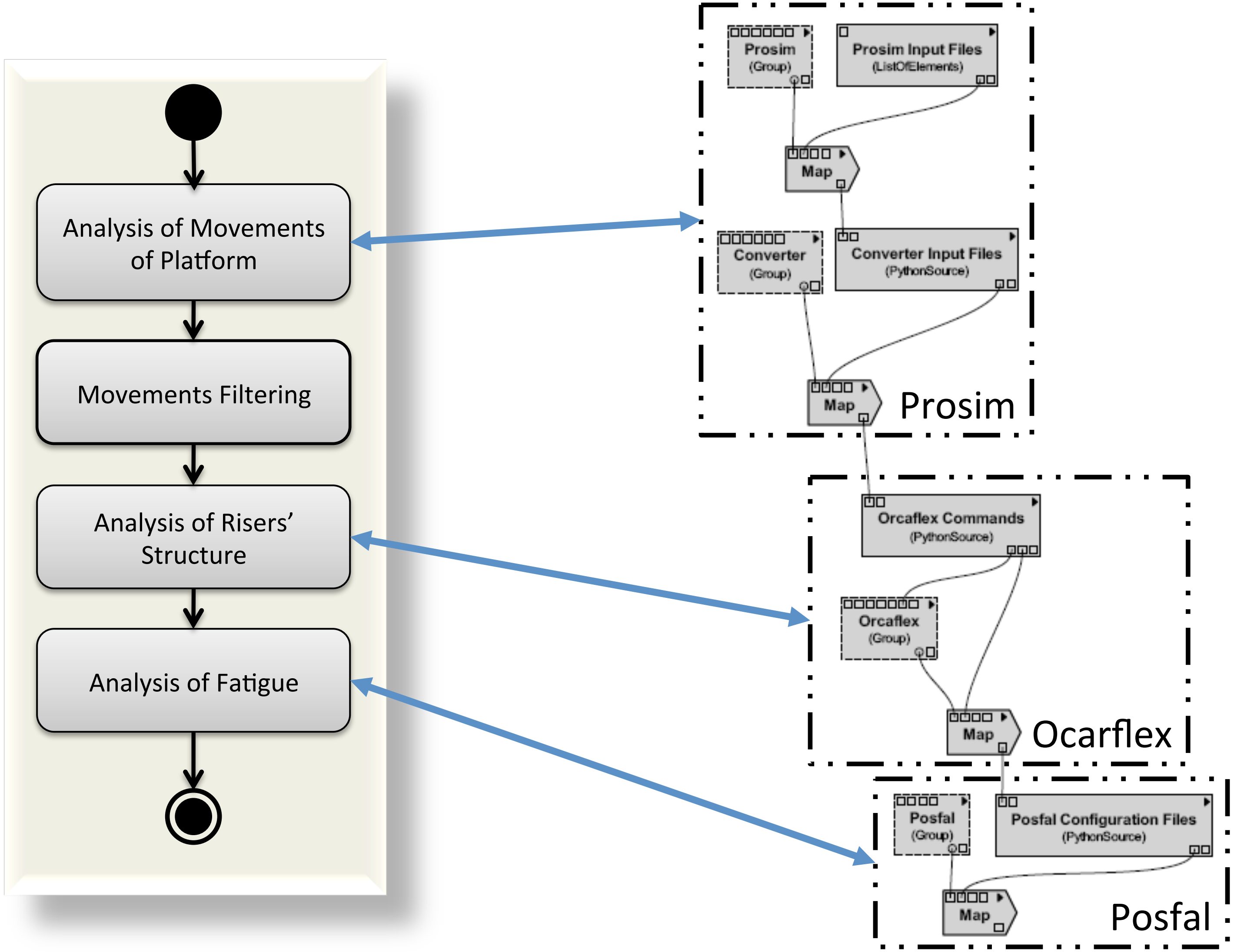
Concrete Workflows for Ultra Deep Water Oil Exploration


Prosim (Group)
Prosim Input Files (ListOfElements)
Map
Converter (Group)
Converter Input Files (PythonSource)
Map
Prosim
TPN (Gr...)
TPN Input Files (ListOfElements)
Map
Converter (Group)
Converter Input Files (PythonSource)
Map
TPN
Orcaflex Commands (PythonSource)
Orcaflex (Group)
Map
Ocarflex
Posinal
Possinal (PythonS...)
Anflex Commands (PythonSource)
Anflex (Group)
Map
Anflex
Posfal (Group)
Posfal Configuration Files (PythonSource)
Map
Posfal
Posfal (Group)
Posfal Configuration Files (PythonSource)
Map
Posfal

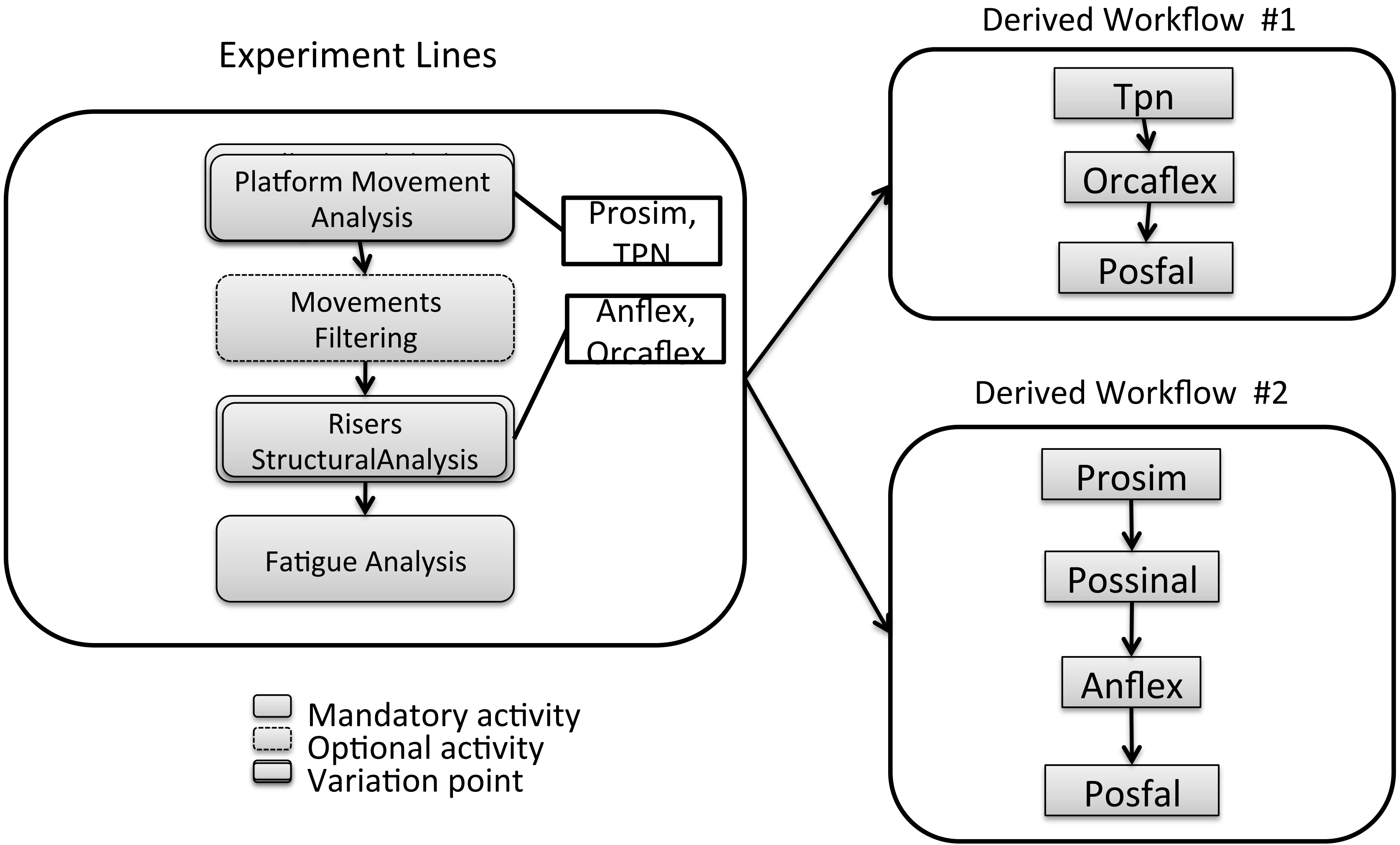
Workflow #1

Workflow #2

Conceptual Workflows and Concrete Workflows Limitations


Analysis of Movements of Platform
Movements Filtering
Analysis of Risers' Structure
Analysis of Fatigue
Prosim (Group)
Prosim Input Files (ListOfElements)
Map
Converter (Group)
Converter Input Files (PythonSource)
Map
Prosim
Orcaflex Commands (PythonSource)
Orcaflex (Group)
Map
Ocarflex
Posfal (Group)
Posfal Configuration Files (PythonSource)
Map
Posfal

Experiment Lines*- abstract workflow



Experiment Lines

Platform Movement Analysis

Movements Filtering

Risers StructuralAnalysis

Fatigue Analysis

Prosim, TPN

Anflex, Orcaflex

Derived Workflow #1

Tpn

Orcaflex

Posfal

Derived Workflow #2

Prosim

Possinal

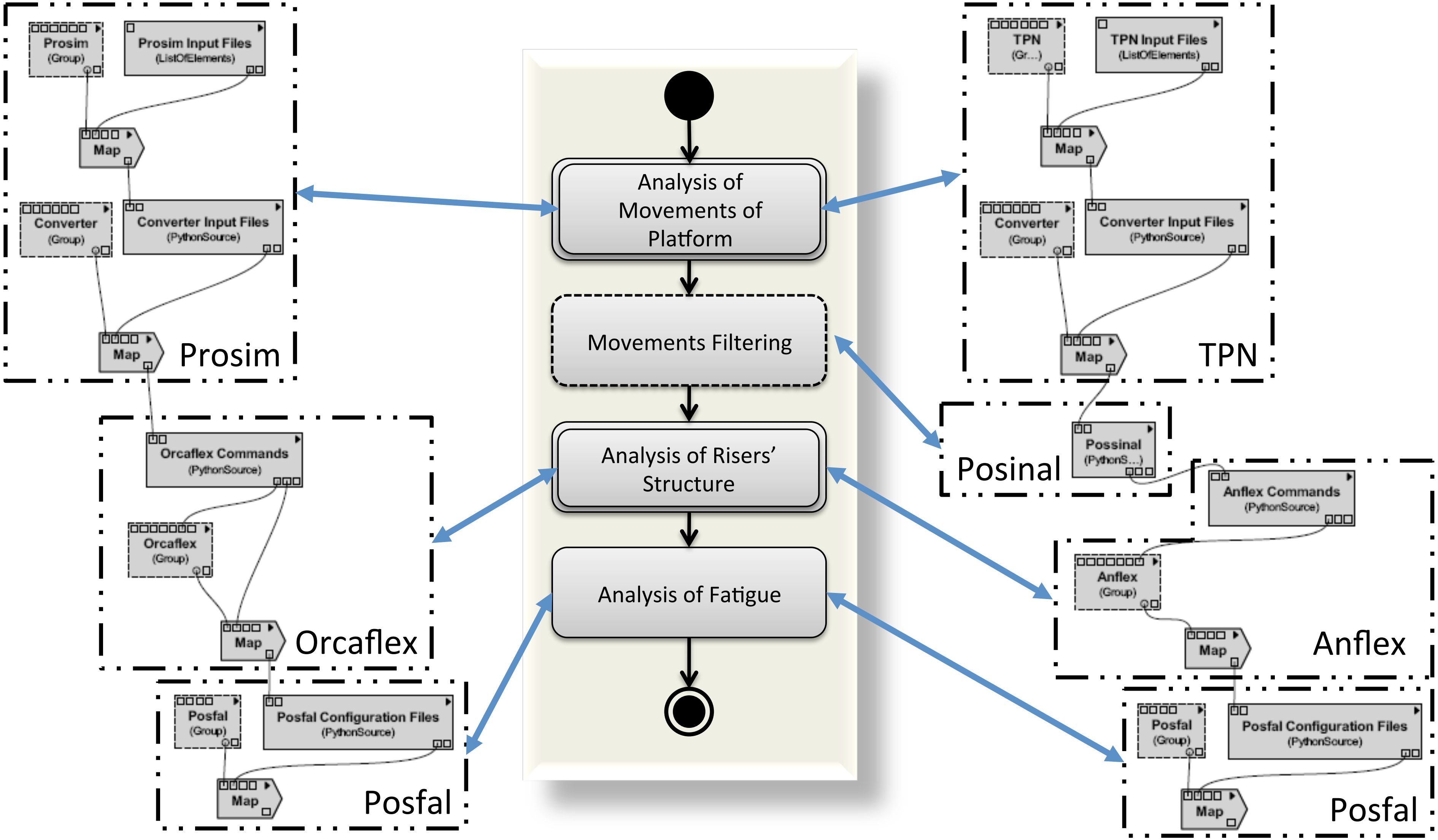
Anflex

Posfal

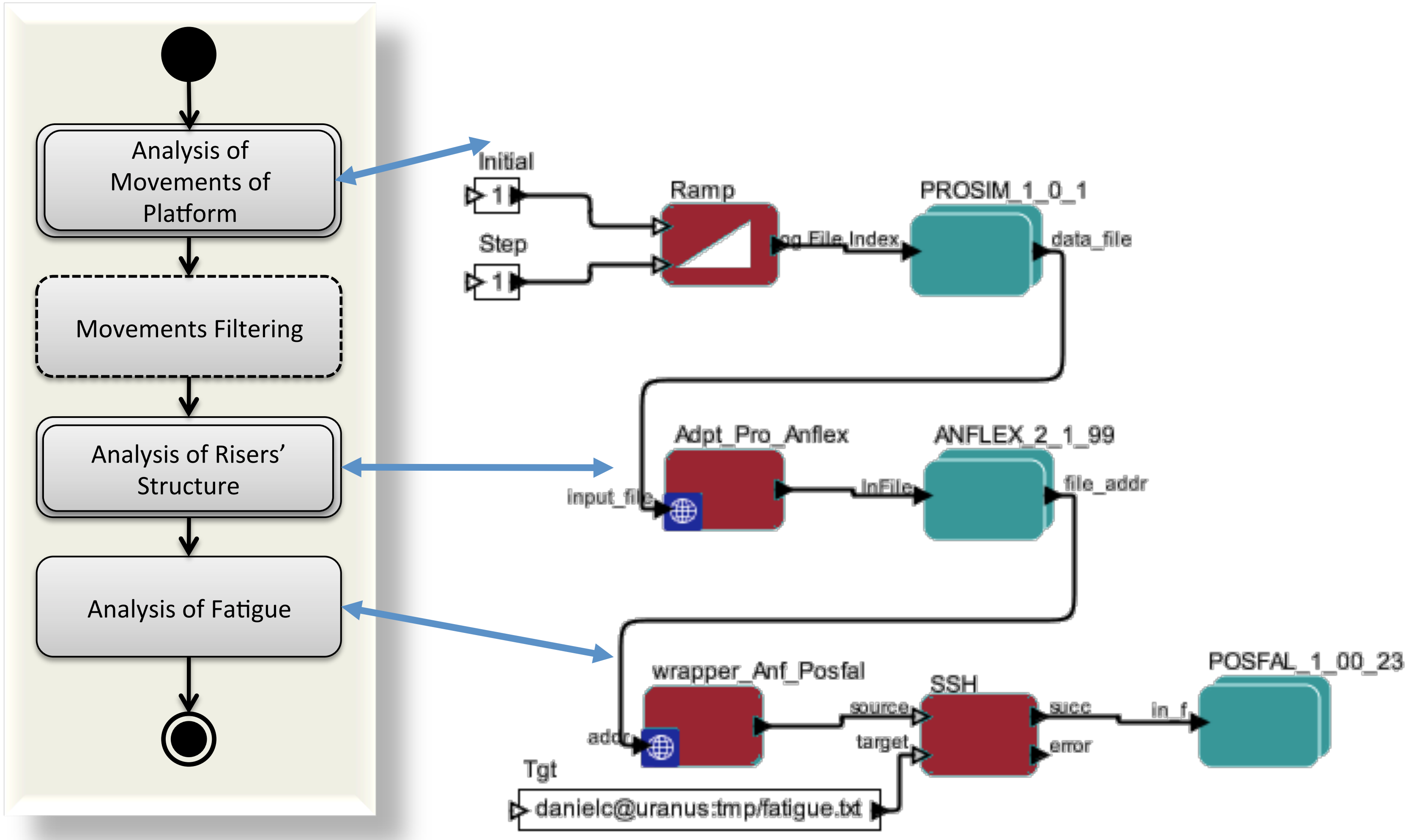
▭ Mandatory activity
⬚ Optional activity
▭ Variation point

*Experiment Line: Software Reuse in Scientific Workflows. SSDBM 2009: 264-272

Workflow Derivation - VisTrails



Workflow Derivation – Kepler



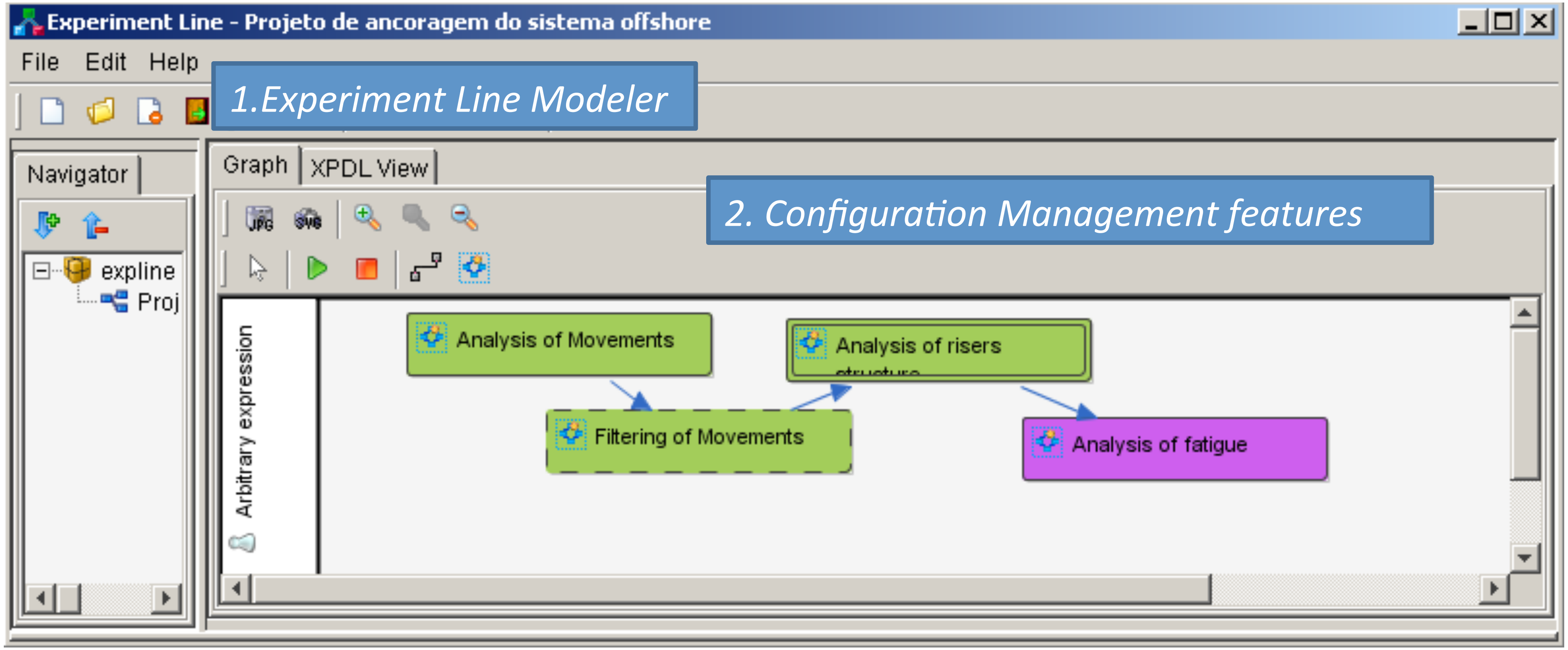
Analysis of Movements of Platform

Movements Filtering

Analysis of Risers' Structure

Analysis of Fatigue

Derivation in GExpLine



1.Experiment Line Modeler

2. Configuration Management features

3. Workflow Importing

4. Workflow derivation

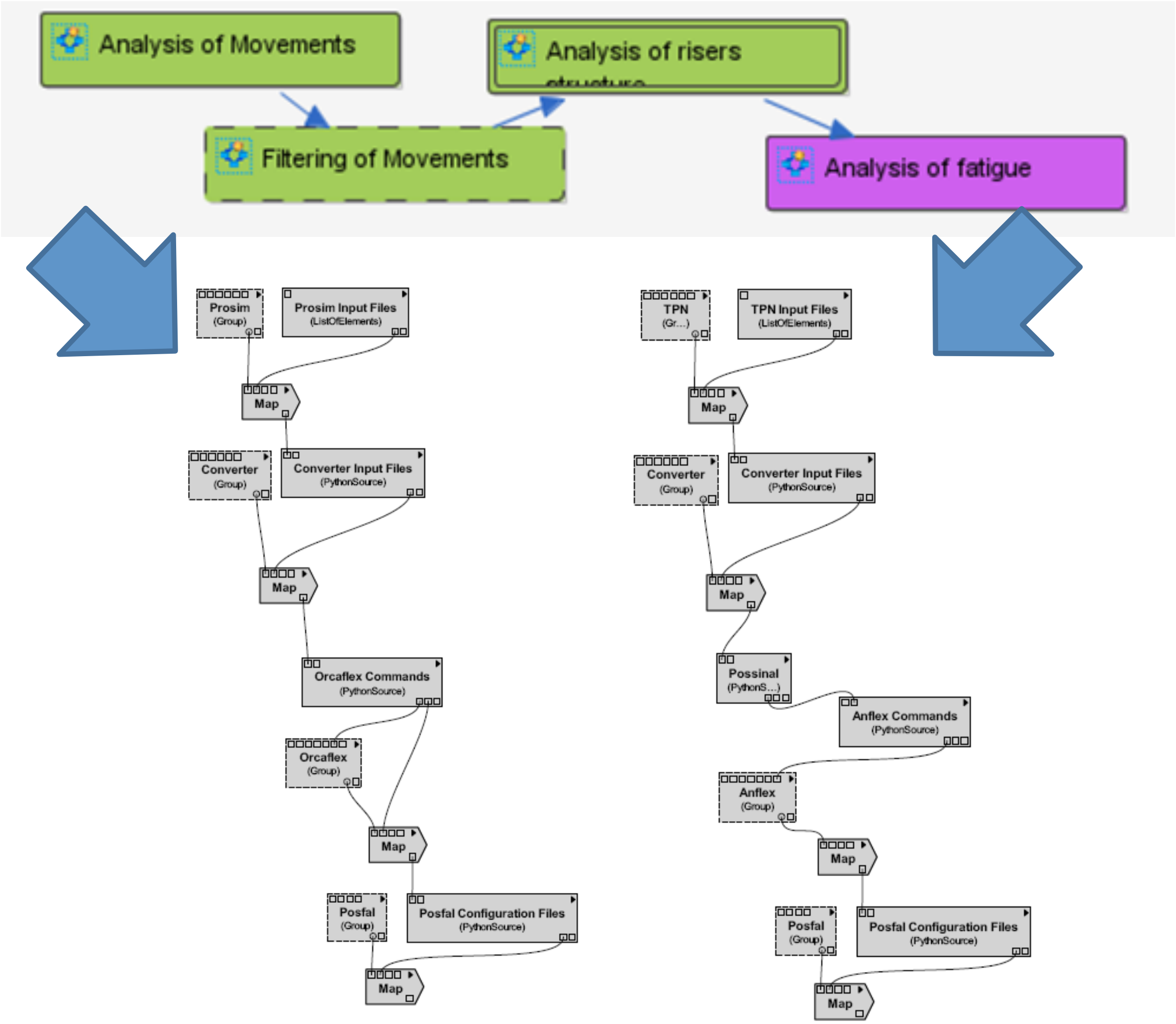
5. Prospective Provenance Querying Support

Derivation Process

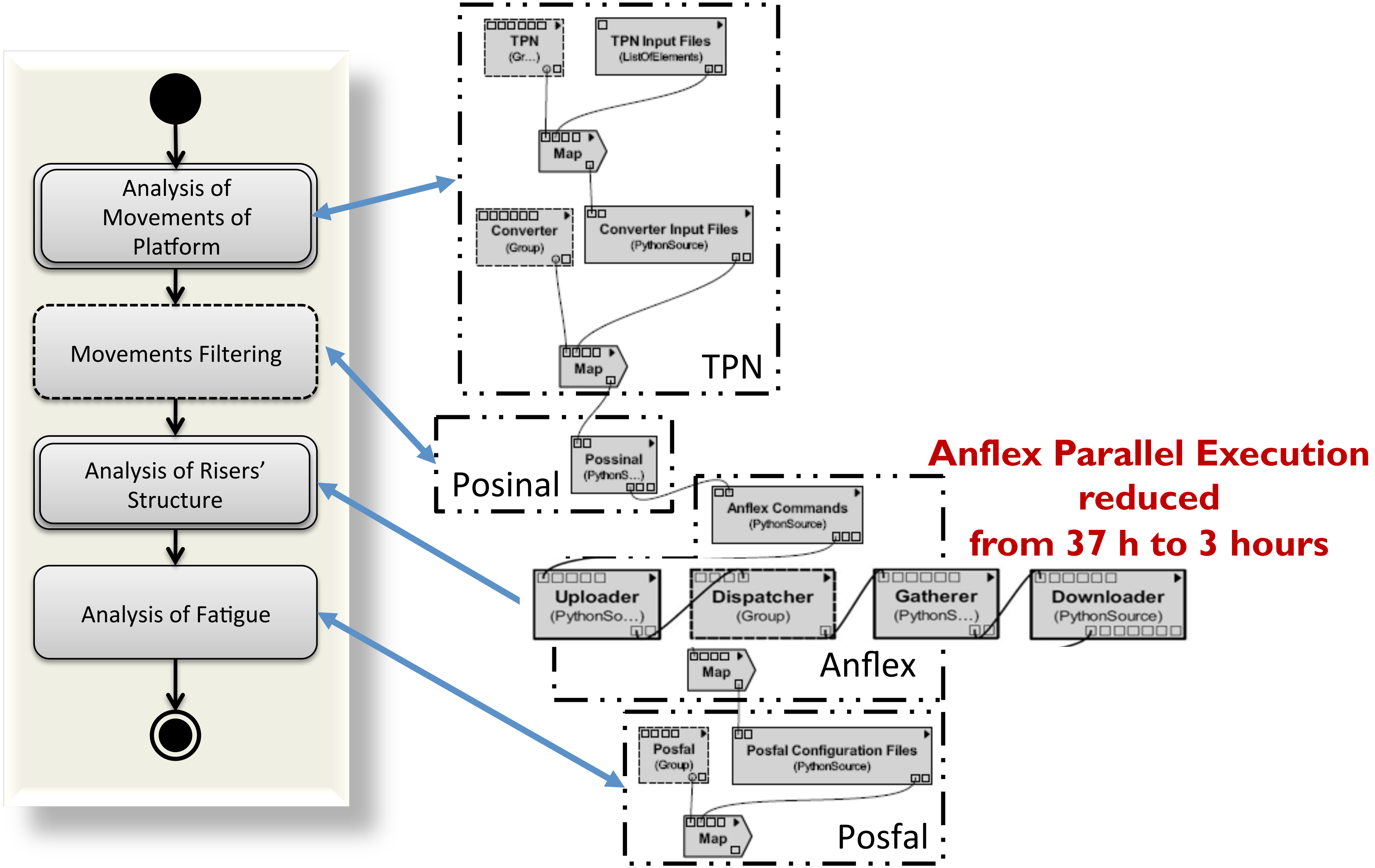
Derive concrete workflows from a conceptual workflow

Derivation information is an important provenance data

It relates all concrete workflows (trials) for a single experiment (conceptual)


Analysis of Movements
Analysis of risers structure
Filtering of Movements
Analysis of fatigue
Prosim (Group)
Prosim Input Files (ListOfElements)
Map
Converter (Group)
Converter Input Files (PythonSource)
Map
Orcaflex Commands (PythonSource)
Orcaflex (Group)
Map
Posfal (Group)
Posfal Configuration Files (PythonSource)
Map
TPN (Gr...)
TPN Input Files (ListOfElements)
Map
Converter (Group)
Converter Input Files (PythonSource)
Map
Possinal (PythonS...)
Anflex Commands (PythonSource)
Anflex (Group)
Map
Posfal (Group)
Posfal Configuration Files (PythonSource)
Map

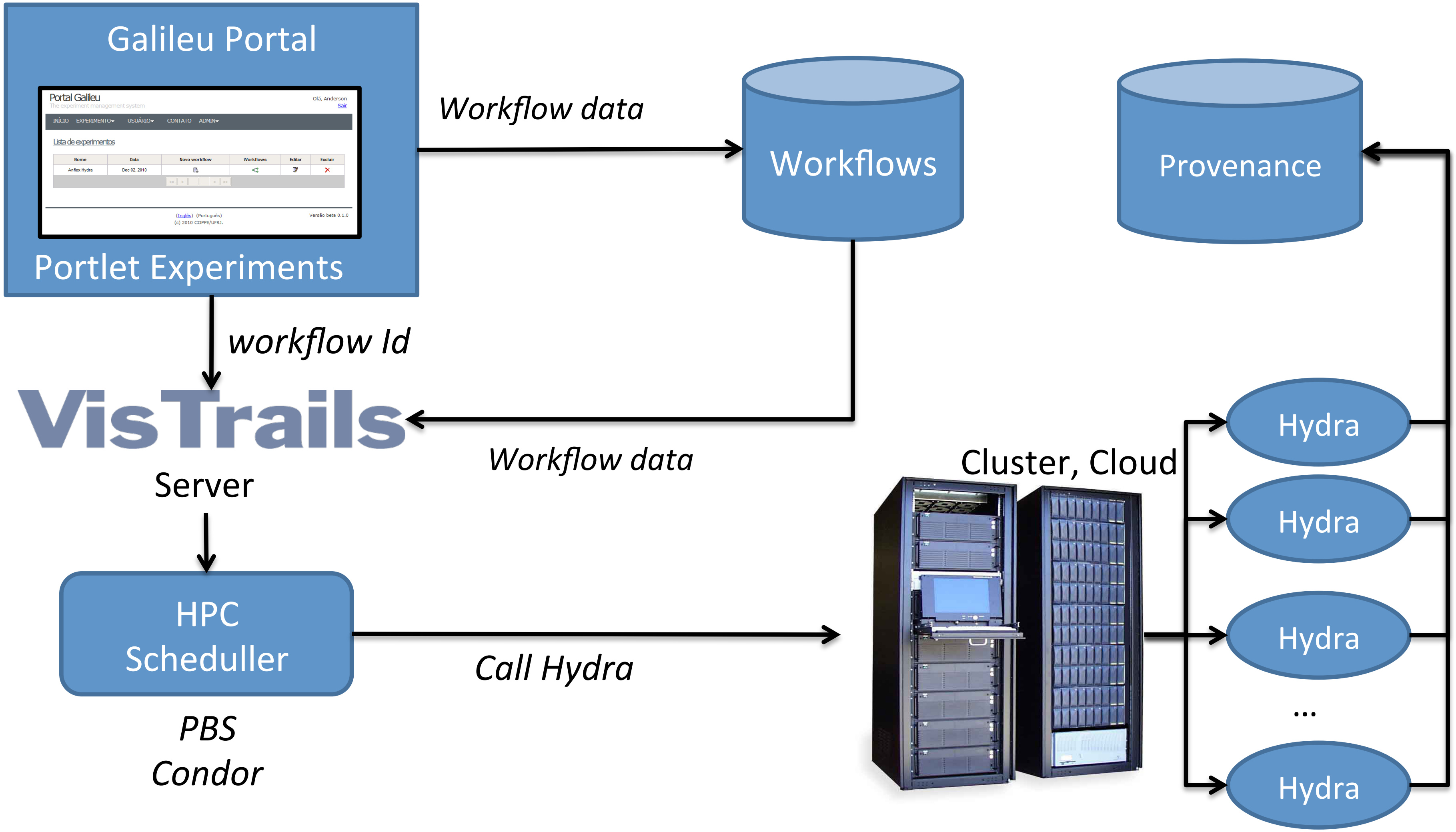
Workflow Derivation – VisTrails and HPC



Analysis of Movements of Platform
Movements Filtering
Analysis of Risers' Structure
Analysis of Fatigue
TPN (Gr...)
TPN Input Files (ListOfElements)
Map
Converter (Group)
Converter Input Files (PythonSource)
Map
TPN
Possinal (PythonS...)
Posinal
Anflex Commands (PythonSource)
Uploader (PythonSo...)
Dispatcher (Group)
Gatherer (PythonS...)
Downloader (PythonSource)
Map
Anflex
Posfal (Group)
Posfal Configuration Files (PythonSource)
Map
Posfal

Anflex Parallel Execution reduced from 37 h to 3 hours

Workflow Execution



Galileu Portal

Portal Galileu

The experiment management system

Olá, Anderson

Sair

Lista de experimentos

Nome	Data	Novo workflow	Workflows	Editar	Excluir
Anflex Hydra	Dec 02, 2010				✗

(Inglês) (Português)

(c) 2010 COPPE/UFRJ.

Versão beta 0.1.0

Portlet Experiments

Issues in distributed provenance

- Provenance integration (local SWfMS and HPC wf execution)
- Provenance gathering in distributed/ heterogeneous environments
- Controlling provenance from parallel execution in distributed environments
- Using provenance for steering activities in distributed environments

Provenance can support analyzing scientific experiments

- Before execution:
  - What programs may be used? Is there any alternative methodology to explore?
  - Is there any dependency between activities? Which activities are mandatory?
- After execution:
  - What were the parameters used in the best result ?
  - What was the scientific workflow used to obtain such result?
  - Where are the output files generated by the distributed activity A using the parameters P?
  - How many times the activity A in version V was used in the experiment E?

all these queries are related to the ability of reproducing and validating a scientific experiment

Provenance Exchange, Integration and Querying

Contributors:

- M. David Allen, Adriane Chapman, Barbara Blaustein, Len Seligman
  [5 Getting It Together: Enabling Multi-organization Provenance Exchange]
- Anderson Marinho, Marta Mattoso, Claudia Werner, Vanessa Braganholo and Leonardo Murta
  [33 Challenges in managing implicit and abstract provenance data: experiences with ProvManager]
- Luiz M. R. Gadelha Jr., Marta Mattoso, Michael Wilde, Ian Foster
  [26 Provenance Query Patterns for Many-Task Scientific Computing]

TaPP '11
3rd USENIX Workshop on the Theory and Practice of Provenance
June 20–21, 2011 | Heraklion, Crete, Greece
usenix
Sponsored by USENIX in cooperation with ACM SIGMOD, ACM SIGPLAN, and the W3C Greece Office

Provenance Exchange, Integration and Querying

Marta Mattoso

Federal University of Rio de Janeiro, Brazil