Scientific Workflow reusing and long term big data preservation

Salima Benbernou

Université Paris Descartes

Salima.benbernou@parisdescartes.fr

Project Mastodons CNRS Predon

Outline

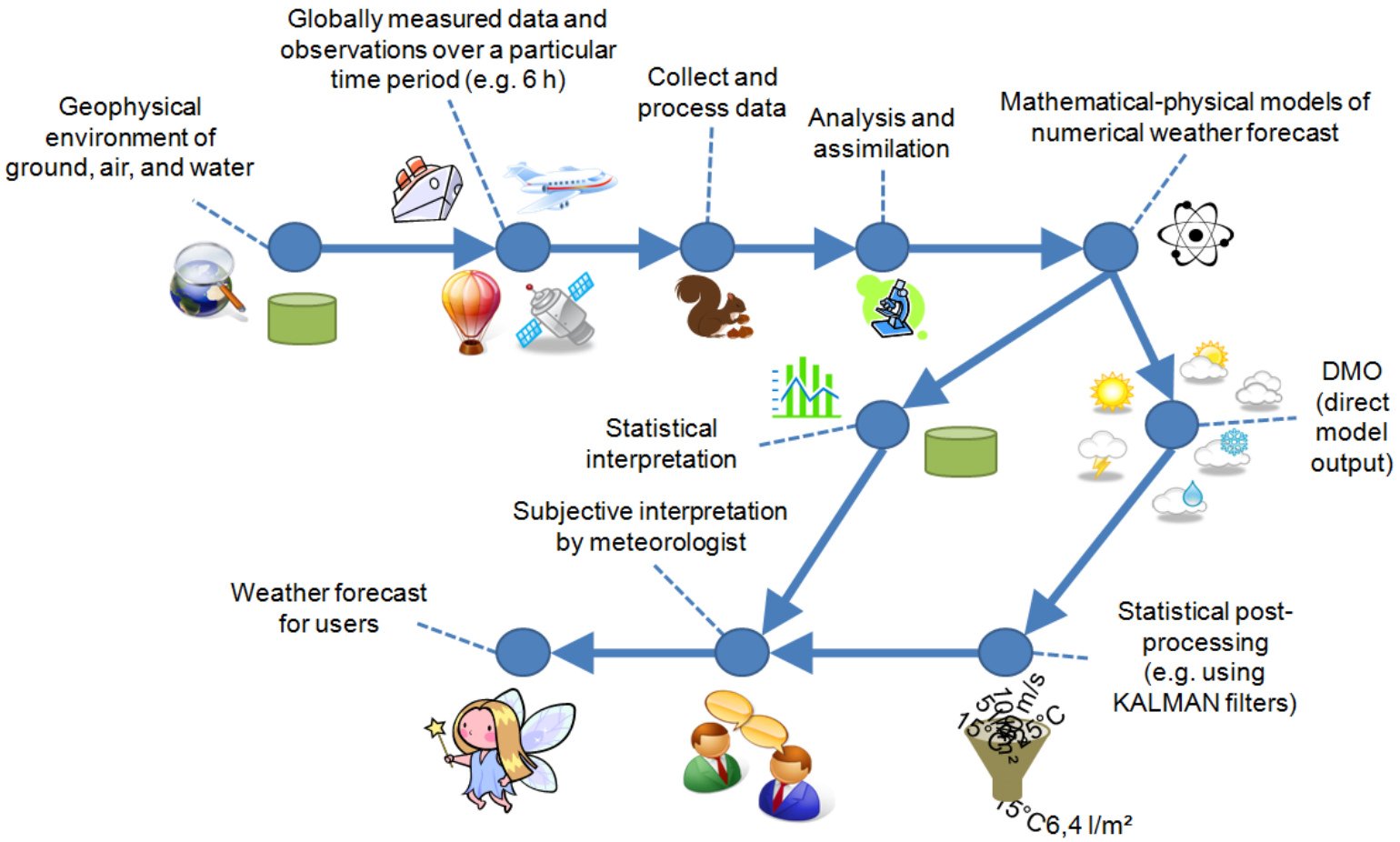
- Scientific workflows
  - Examples
  - Existing systems and limitations
- Using conventionnel workflow technologies in simulation/experiments
  - Introduction
  - Modeling using BPEL
- Reusing in scientific workflow
  - Fragment reusing
  - Privacy aware provenance

What are scientific workflows?

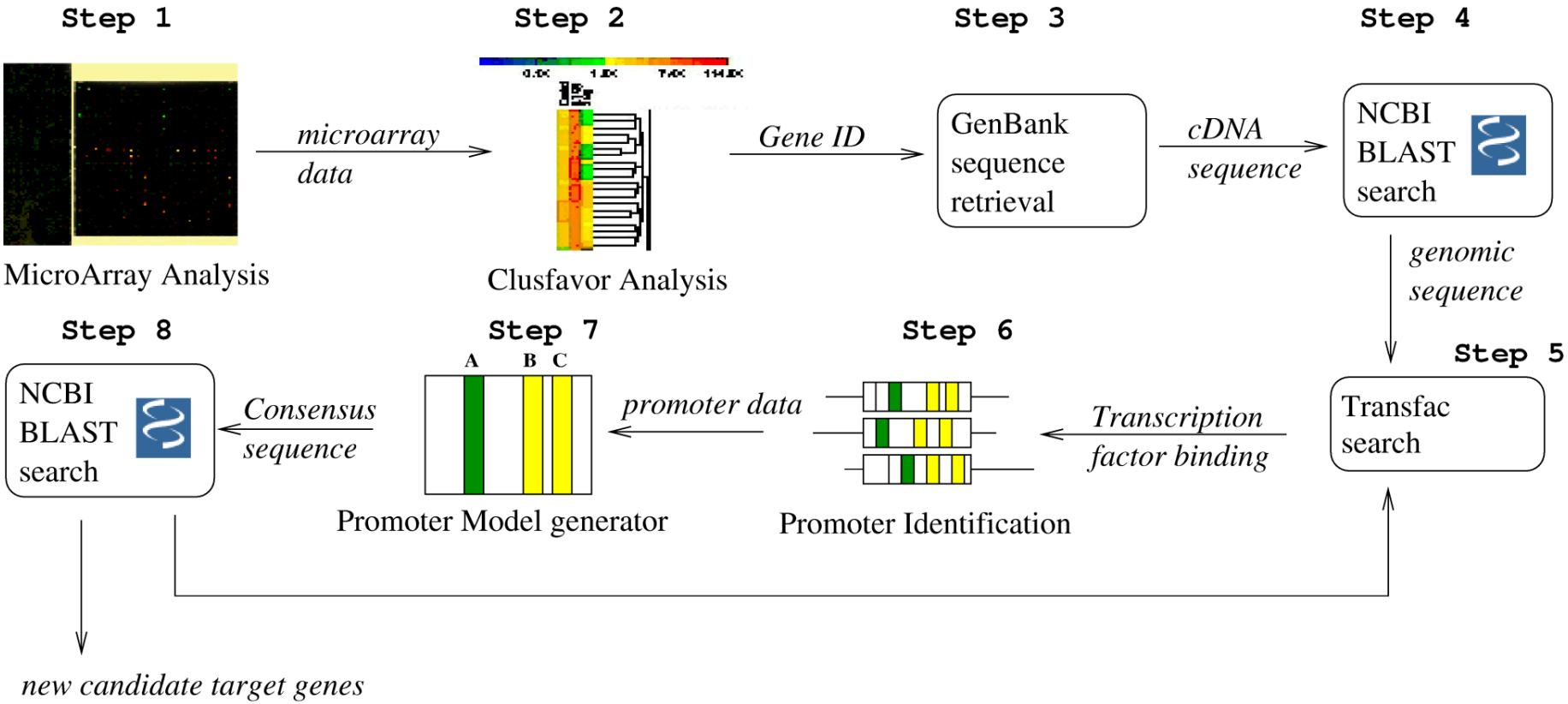
- Scientific experiments/computations/simulation modeled and executed as workflows called scientific workflow (SWf).

- Deal with intensive data, are long running, data driven, can integrate multiple data sources (i.e. sensors)

SWF Examples

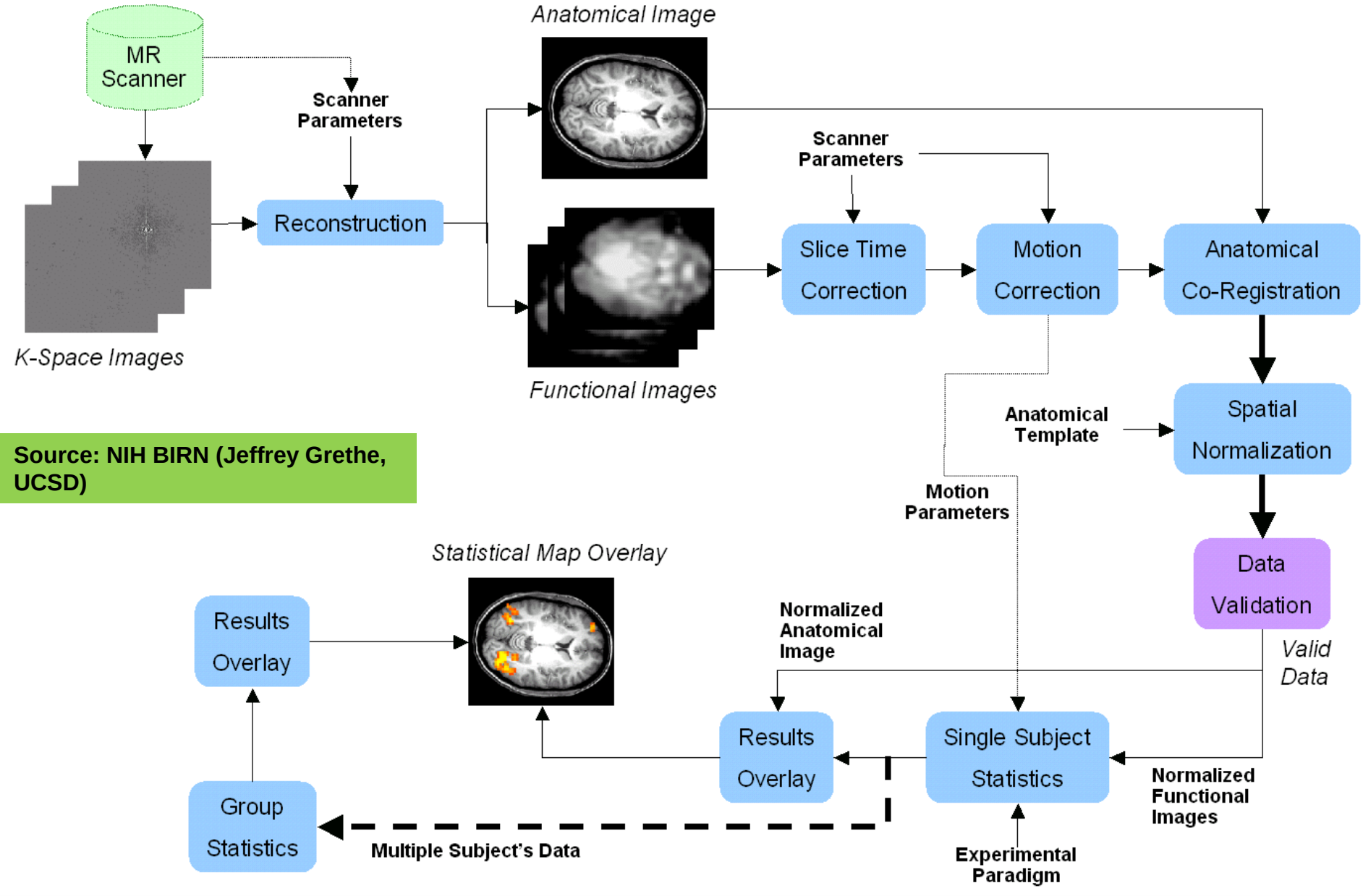


SWF Examples



Source: Matt Coleman (LLNL)

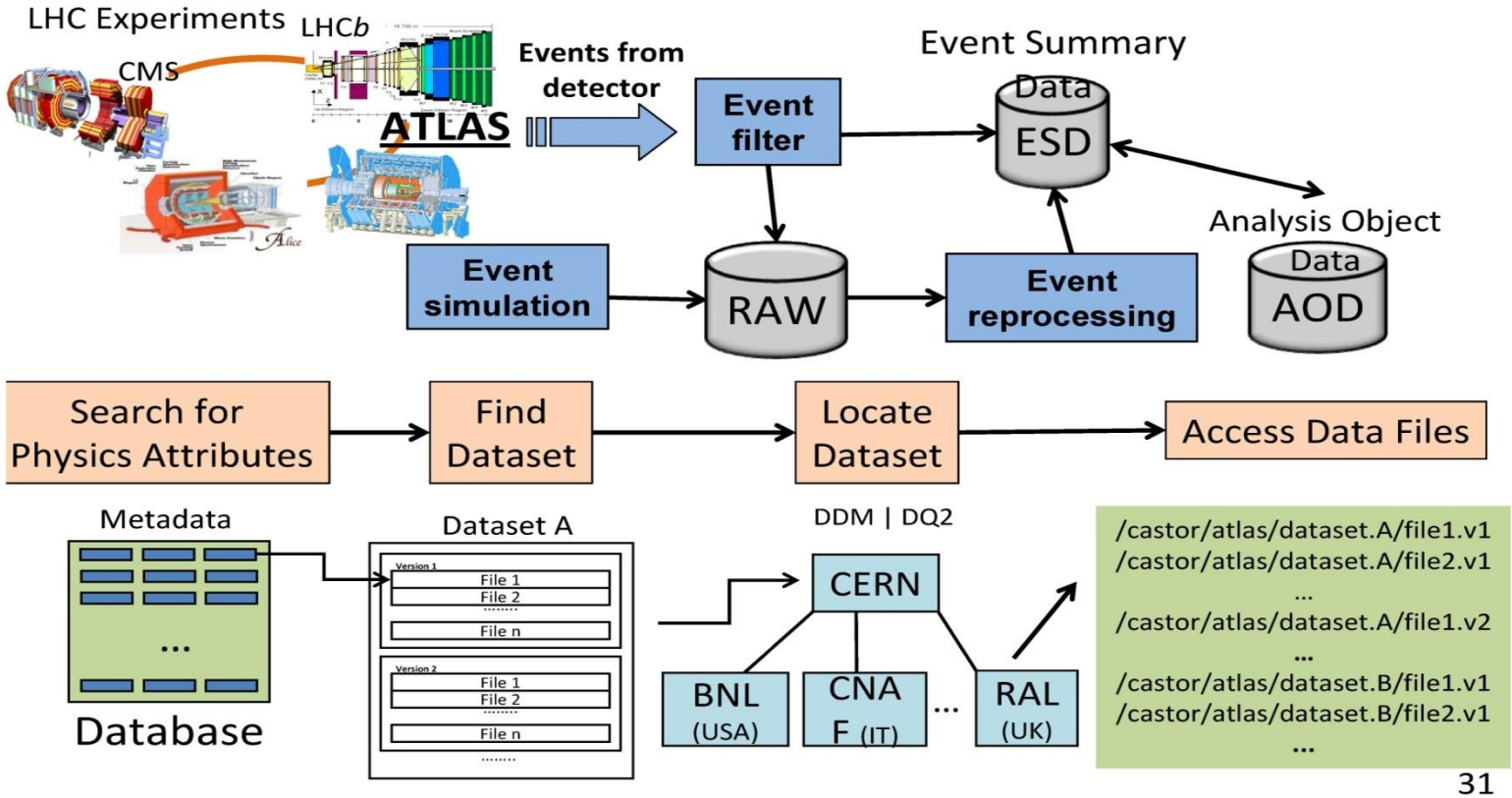
Functional MRI Analysis Workflow



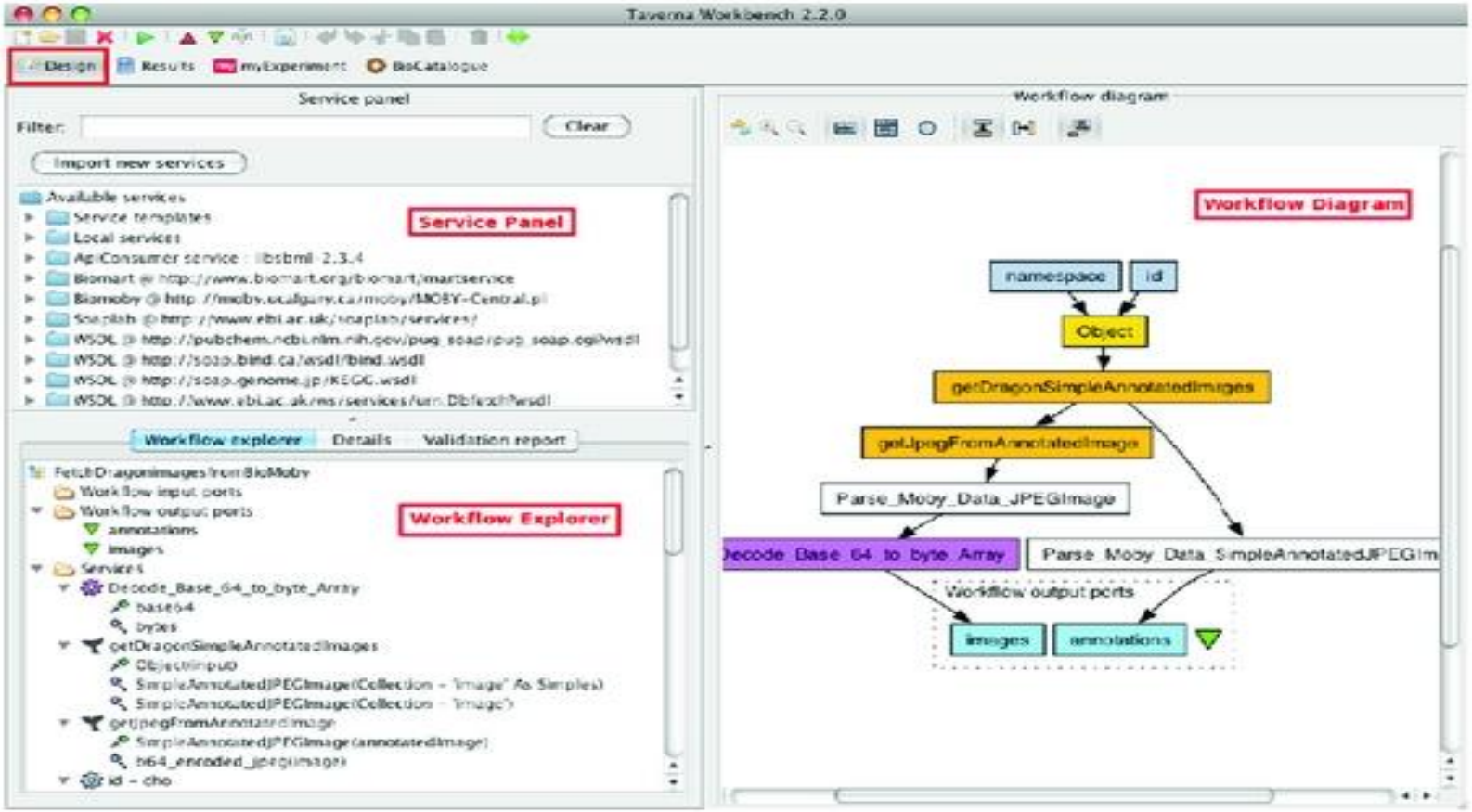
Source: NIH BIRN (Jeffrey Grethe, UCSD)

SWF Examples (cont)

ATLAS experiment (simplified)



SWF Examples (cont)



A snapshot of the Taverna Workbench.

Scientific workflow systems

- Workflow are already used in e-science
- This is not always the conventional workflow technology
- Some workflow systems in e-science: Kepler, Taverna, Pegasus, Trident, Simulink, Karajan …

- To be improved
  - Robustness, fault handling
  - Flexibility and adaptability
  - Reusability
  - Scalability
  - Interaction with users, user-friendliness of tools
  - science skills required from scientist
  - No generic approach
  - Domain specific solutions (in term of modeling and execution)

Scientific workflow systems

- Data-driven applications are more and more developed in science to exploit the large amount of digital data today available
- Adequate workflow composition mechanisms are needed to support the complex workflow management process including workflow creation, workflow reuse, and modifications made to the workflow over time.
- Use conventional technologies **(Business processes)**

Business workflows (i.e, BPEL)

- independent of the application domain, can be used for every type of scenario
- The concept of workflow models and instances is inherently capable of enabling parameter sweeps.
- Asynchronous messaging features are predestinated for non-blocking invocation of long running scientific computations
- Business workflows are usually based on agreed-upon standards for workflow modeling and execution as well as for integration technologies.
- facilitate collaboration between scientists (e.g., with the help of Web services).
- Services computing technology enables scientists to expose data and computational resources wrapped as publicly accessible Web services

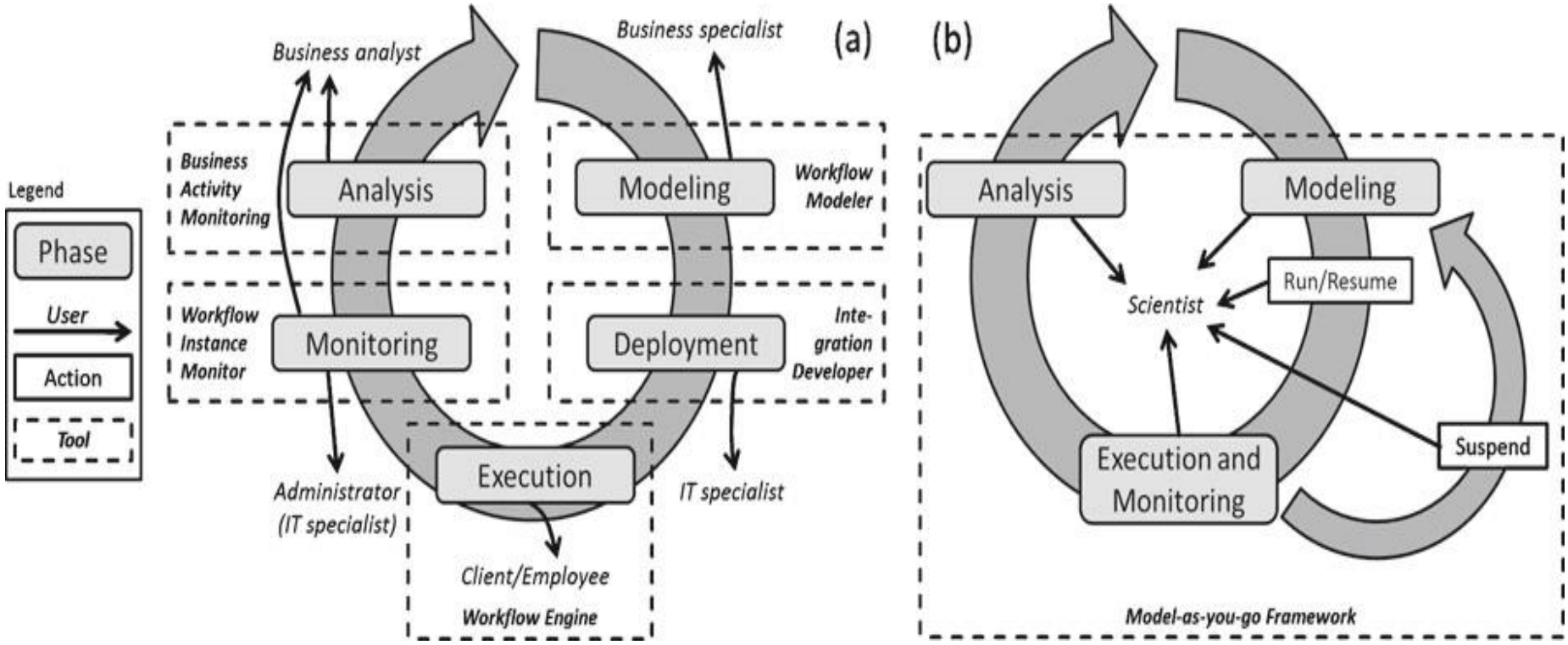
Scientific workflows vs. Business workflows

- Scientific "Workflows"
  - Dataflow and data transformations
  - Data problems: volume, complexity, heterogeneity
  - Grid-aspects
    - Distributed computation
    - Distributed data
  - User-interactions/WF steering
  - Data, tool, and analysis integration

  ➔ Dataflow and control-flow are ***married!***
- Business Workflows
  - Process composition
  - Tasks, documents, etc. undergo modifications (e.g., flight reservation from *reserved* to *ticketed*), but modified WF objects still identifiable throughout
  - Complex control flow, task-oriented: *travel reservations; credit approval*

  ➔ Dataflow and control-flow are ***divorced!***

Business and scientific Lifecycles



Scientific workflow limitations

- Scientific workflow life cycle: scientits'perspective
  - Reflects how scientits actually work-trial and error fashion
  - Hidden technical details
  - « no » deployment phase
  - Operations to control workflow execution
  - Monitoring is the visualisation of the results only

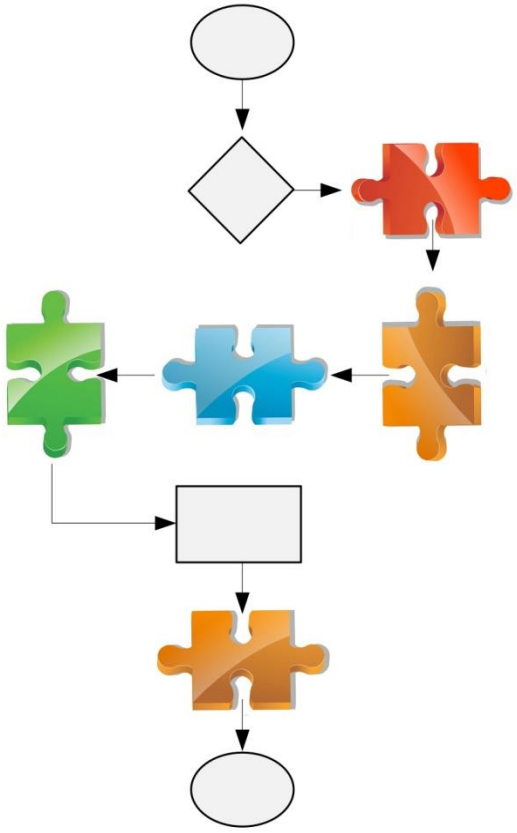
Scientific workflow and the scalability

- Service-Oriented Workflows on Cloud Infrastructures for **reusing**.
- The service-oriented paradigm will allow **large-scale distributed workflows** to be run on heterogeneous platforms and the integration of workflow elements developed by using different programming languages. **Web and Cloud services** are a paradigm that can help to handle workflow interoperability,

BPaas vs SPaaS

- Business process vs swf outsourcing to take advantage of the Cloud computing model.
- Reusing process fragments to develop process-based service compositions and adapt the new swf according to the scientists (reusing a partial differential equation program)
- privacy risks aware.
- **Sharing scientific process fragments and hide provenance.**
- Many works studied data provenance but not hide provenance.
- Formal model of BPaaS vs Scientific Process as a Service (Icloud@vldb2012)

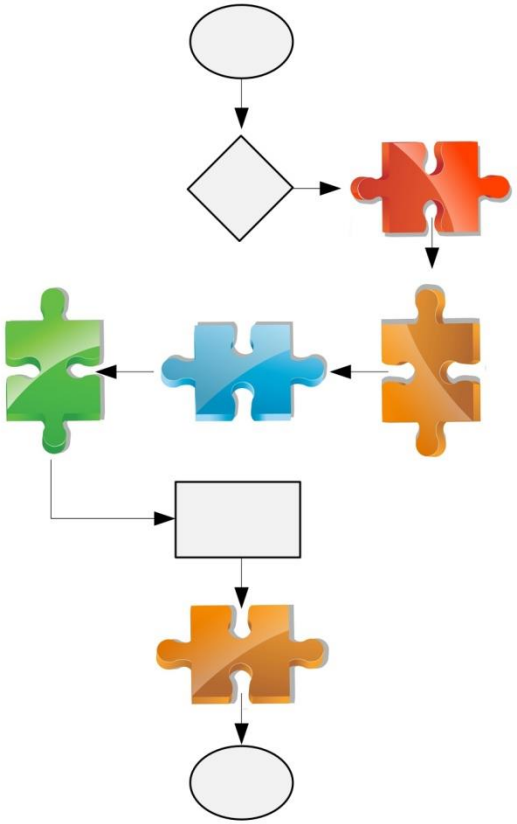
Decomposition of business process vs scientific process



Hospital Business Process

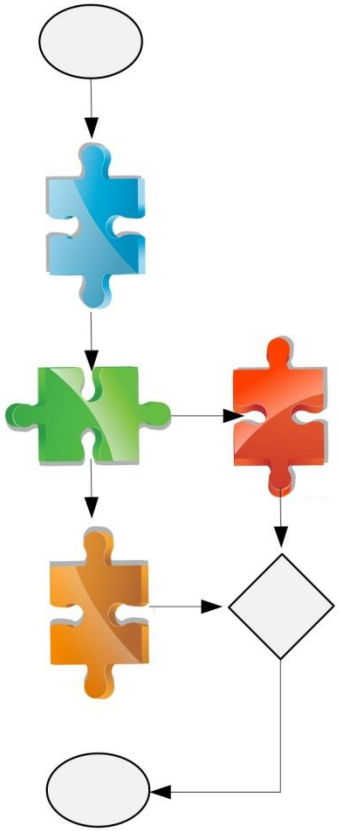
1. A business process is decomposed into a set of process fragments suitable for re-use.

Identification of fragments



1. A business process is decomposed into a set of process fragments suitable for re-use.

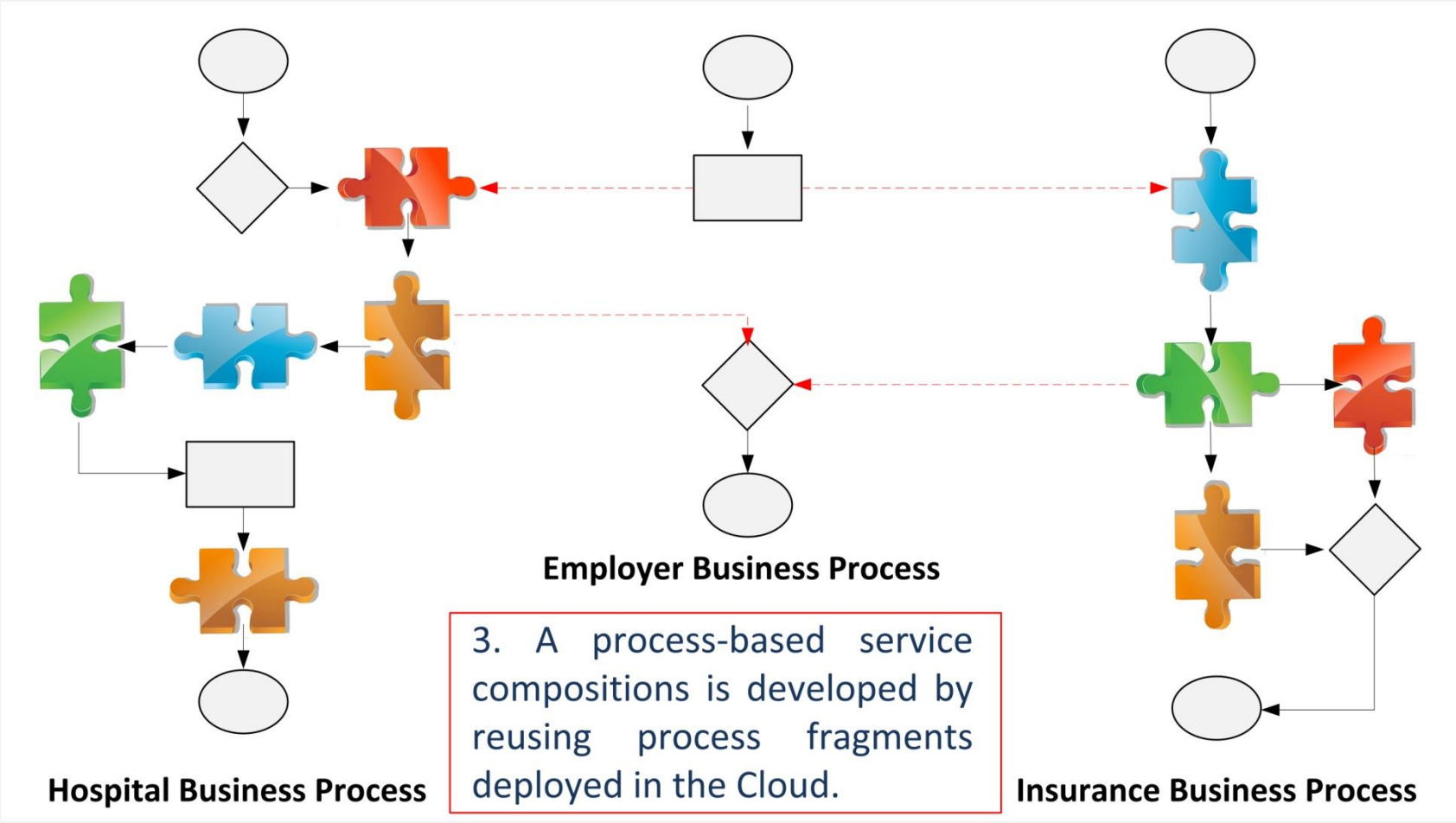
2. Each process fragment is identified and made available to be reused as a Web Service.



Hospital Business Process

Insurance Business Process

Development of processes



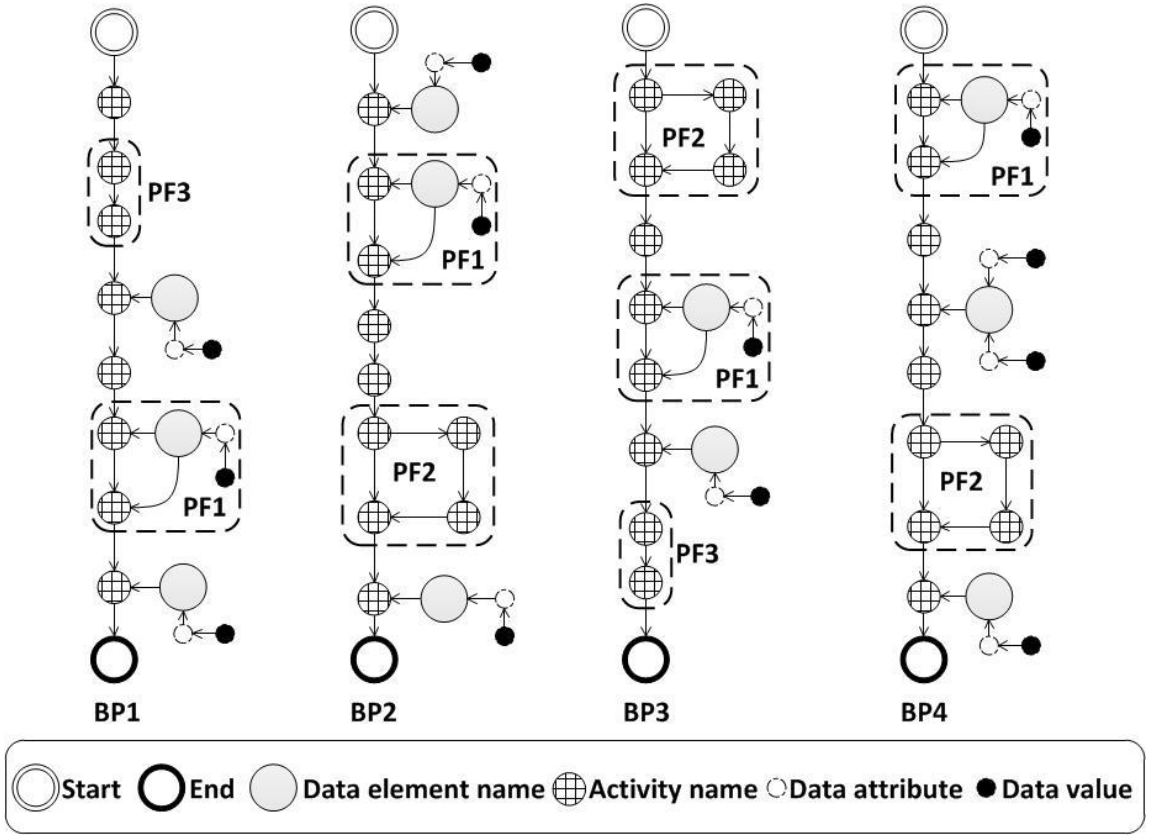
Provenance and privacy in SPaaS

An adversary (a curious) **can discover the provenance** of the reused process fragments.

Can infer connections between end-users and scientists that outsource fragments to the Cloud.

✓**No related work!**

Formal model



- **Business Process:** business graph. [Beeri,VLDB'06]
- ✓ **Process Fragment:** business subgraph.
- ✓ **BPaaS:** a finite set of business processes.
- ✓ **Reusing Function.**

Anonymous Views

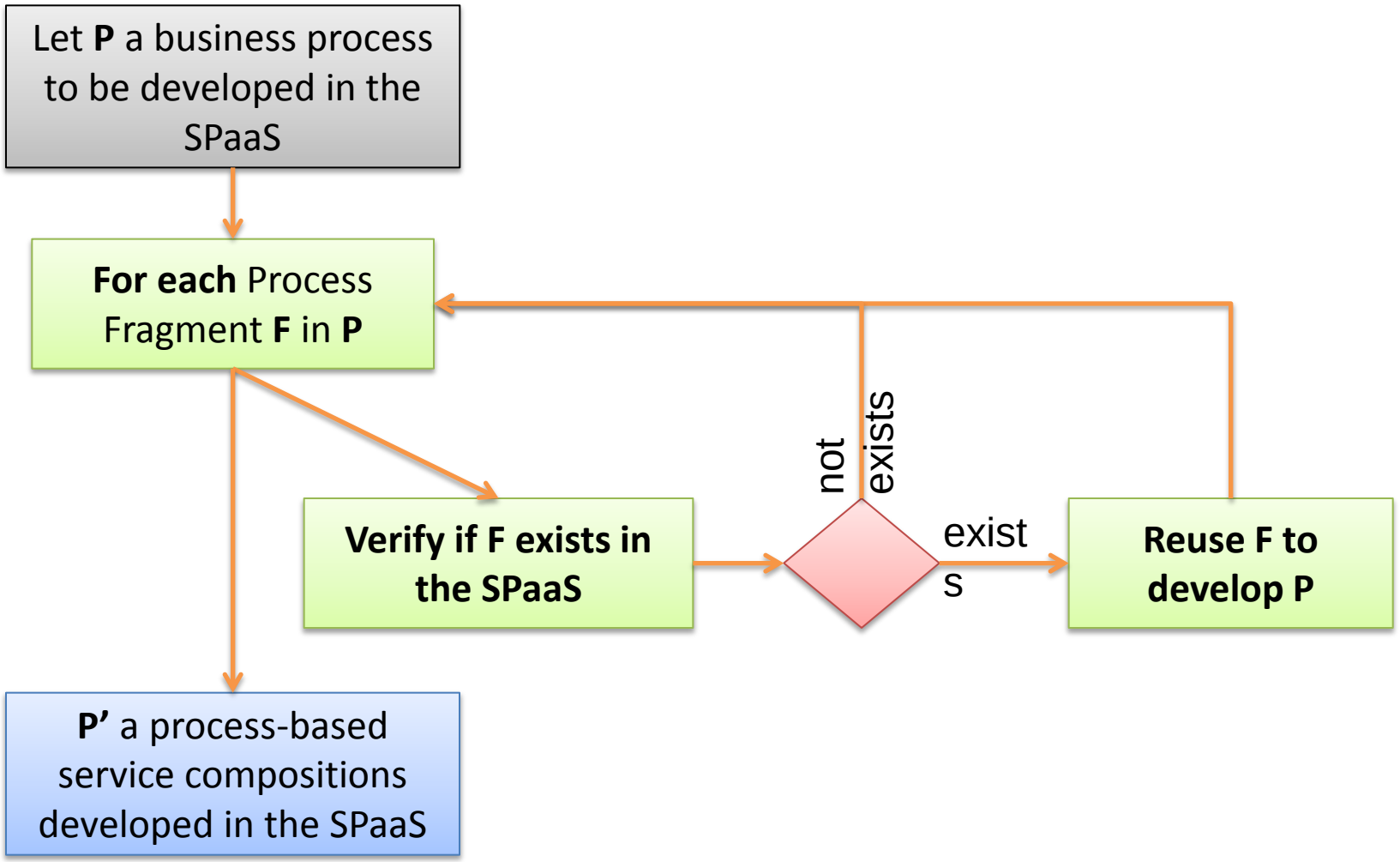
- **View on SPaaS:**

  A set of process fragments having the same objective (called clones).

- **Anonymous View on SPaaS:**

  View on SPaaS having at most K clones.

- **Objective : Make it hard for an adversary to know the provenance of a reused process fragment. (**Anonyfrag)


Let **P** a business process to be developed in the SPaaS
For each Process Fragment **F** in **P**
Verify if F exists in the SPaaS
not exists
exist s
Reuse F to develop P
P' a process-based service compositions developed in the SPaaS


MERCI!
THANK YOU!
FRAPAR.

Workshop organisation

1st Workshop on LOng term Preservation for big Scientific data (LOPS) to be held in conjunction with ICDE 2014, Mach 31-April 4, Chicago , IL, USA

Lipade.math-info.univ-paris5.fr/lops/