UNIVERSITÉ PARIS DESCARTES

LIPADE
Laboratoire d'Informatique PAris DEscartes

Scientific Workflow reusing and long term big data preservation

Salima Benbernou

Université Paris Descartes

Salima.benbernou@parisdescartes.fr

Project Mastodons CNRS Predon

Outline

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

Square Predict

Prédiction financière:

Gestion de données assurances et open data

Square Predict

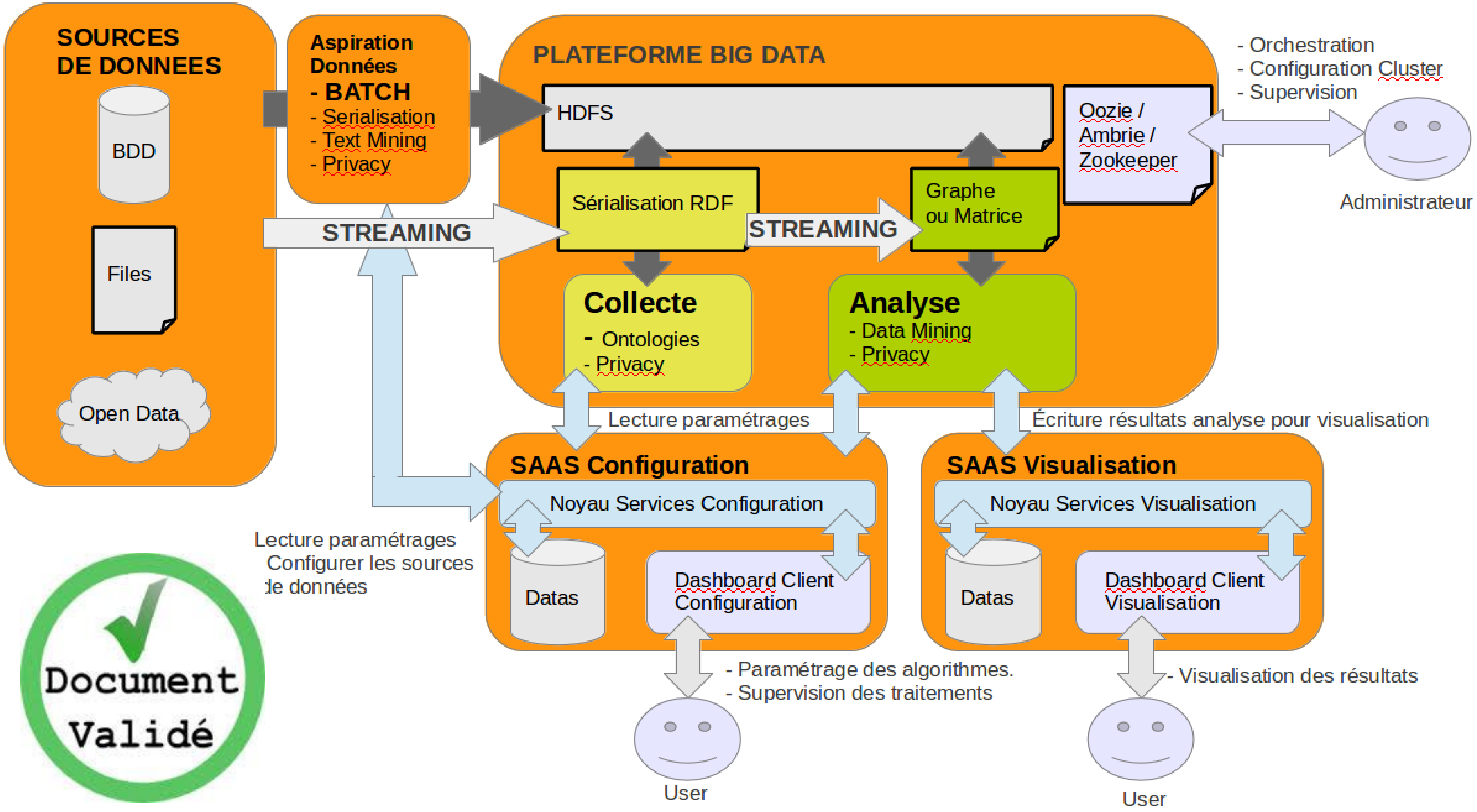
- Collecte de données sur Teralab
- Fusion de données sémantiques (RDFs):

  -Evaluation de requêtes par réecriture distribuée sur Spark

  -Requêtes virtuelles

  -réutilisation des données pour d'autres requêtes
- Qualité de données:inconsistance, l'incertitude
- Prendre les Données en streaming
- Clustering en streaming sur Spark
- Visualisation

SP3.1 – Architecture fonctionnelle



SOURCES DE DONNEES

BDD

Files

Open Data

Aspiration Données

- BATCH
- Serialisation
- Text Mining
- Privacy

STREAMING

PLATEFORME BIG DATA

HDFS

Sérialisation RDF

STREAMING

Graphe ou Matrice

Collecte

- Ontologies
- Privacy

Analyse

- Data Mining
- Privacy

Oozie / Ambrie / Zookeeper

- Orchestration
- Configuration Cluster
- Supervision

Administrateur

Lecture paramétrages

Écriture résultats analyse pour visualisation

SAAS Configuration

Noyau Services Configuration

Datas

Dashboard Client Configuration

Lecture paramétrages
Configurer les sources
de données

- Paramétrage des algorithmes.
- Supervision des traitements

User

SAAS Visualisation

Noyau Services Visualisation

Datas

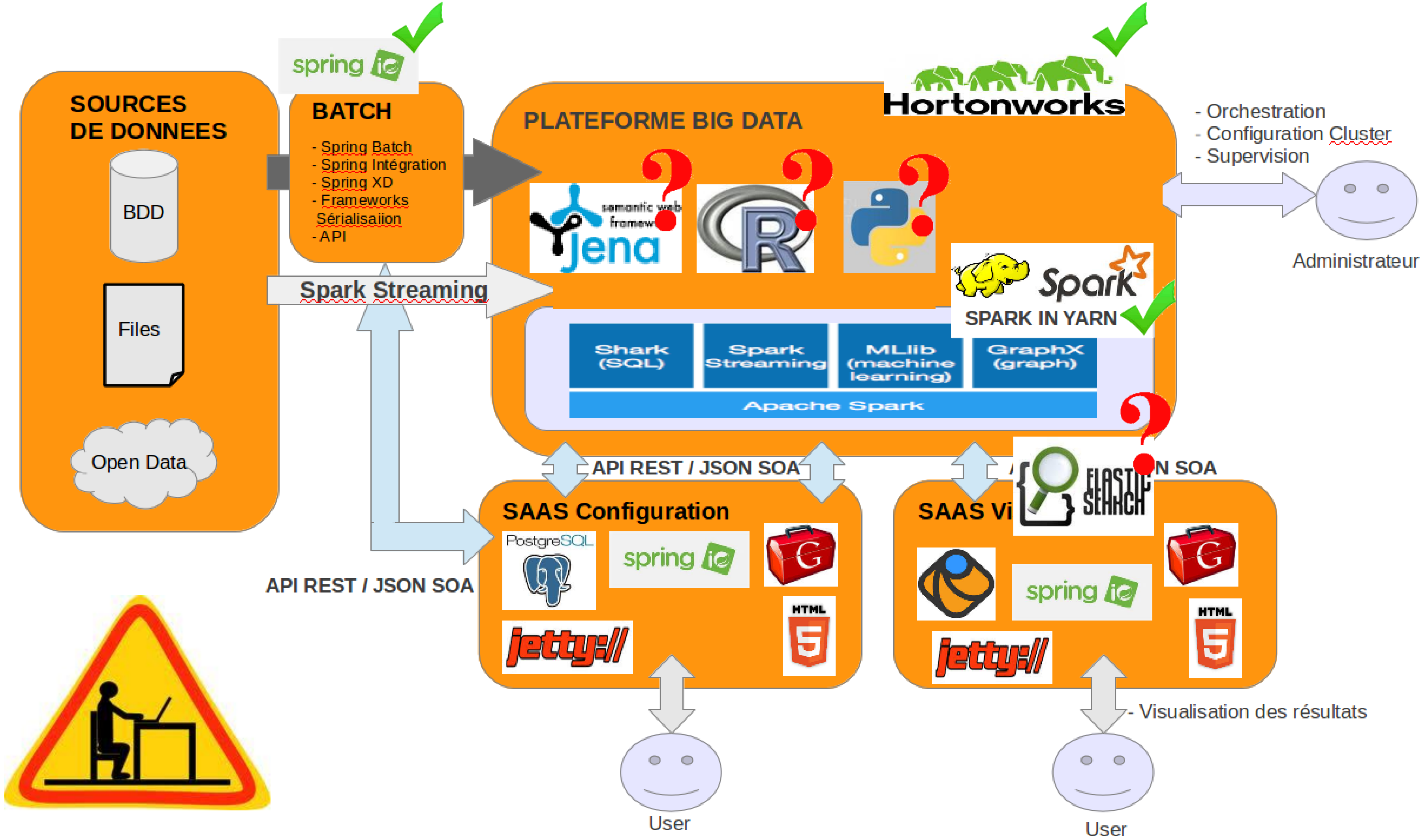
Dashboard Client Visualisation

- Visualisation des résultats

User

Document Validé

SP3.1 – Architecture technique



SOURCES DE DONNEES

BDD

Files

Open Data

BATCH

- Spring Batch
- Spring Intégration
- Spring XD
- Frameworks Sérialisaiion
- API

Spark Streaming

PLATEFORME BIG DATA

SPARK IN YARN

Shark (SQL)

Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

- Orchestration
- Configuration Cluster
- Supervision

Administrateur

API REST / JSON SOA

API REST / JSON SOA

SAAS Configuration

SAAS Vi

User

- Visualisation des résultats

User

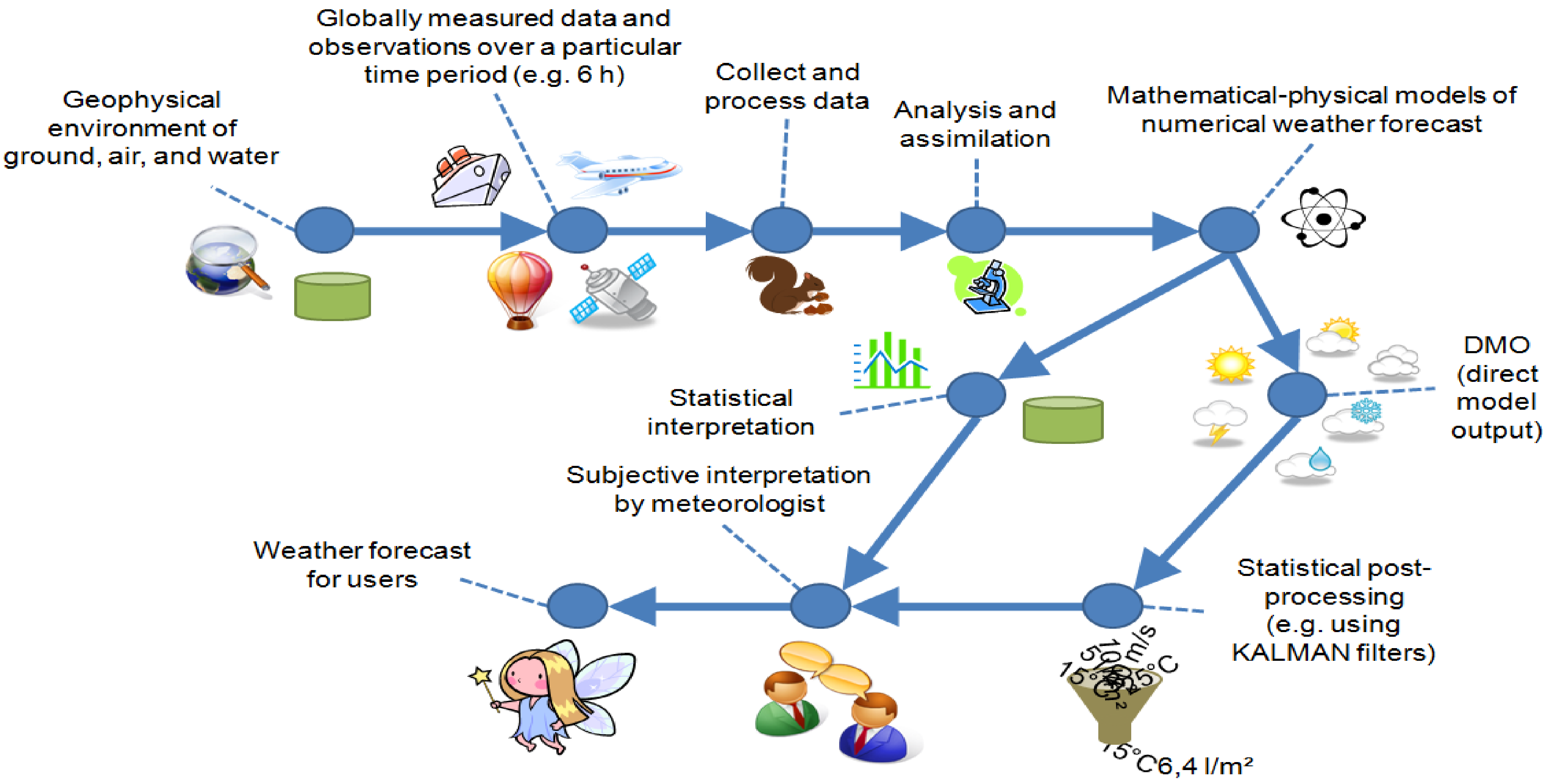
Scientific workflow

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



Geophysical environment of ground, air, and water

Globally measured data and observations over a particular time period (e.g. 6 h)

Collect and process data

Analysis and assimilation

Mathematical-physical models of numerical weather forecast

DMO (direct model output)

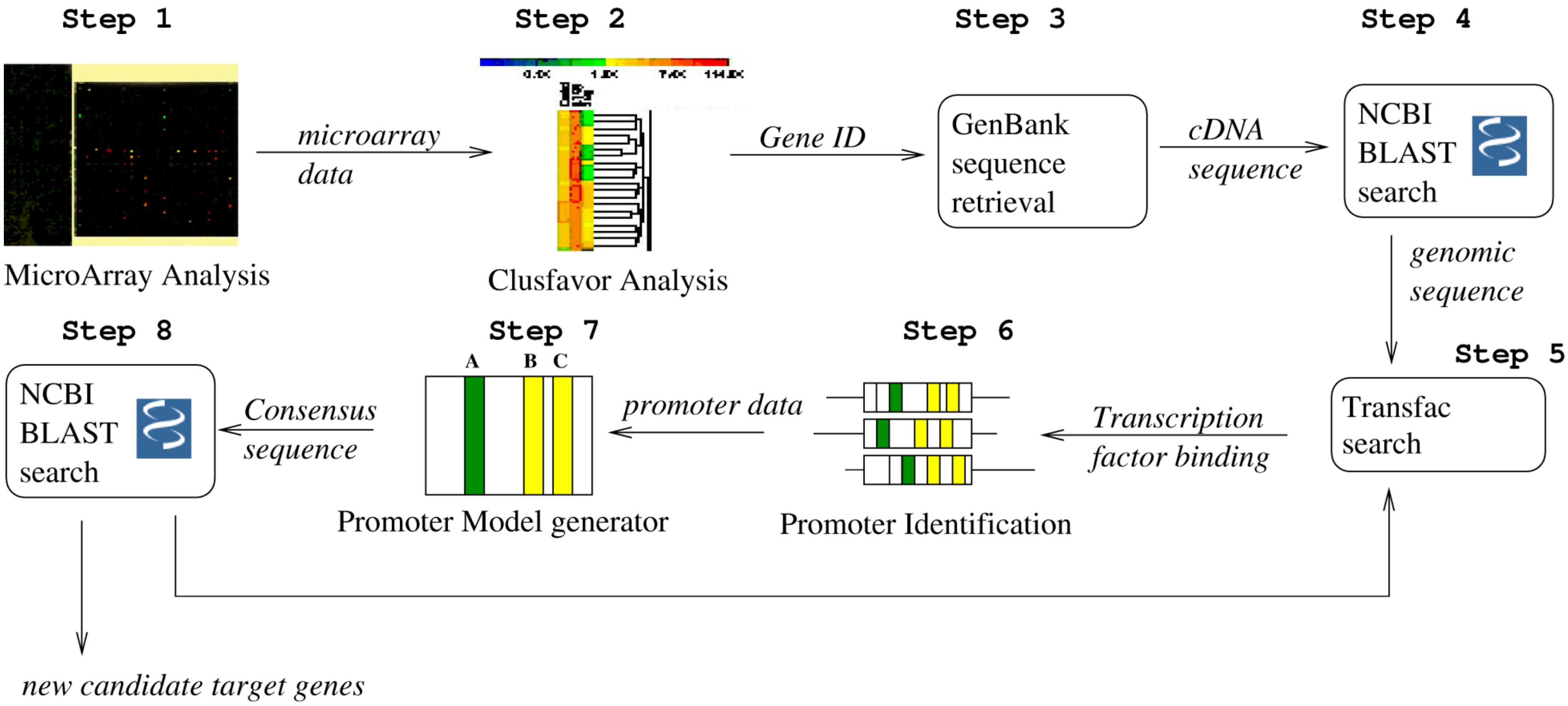
Statistical interpretation

Subjective interpretation by meteorologist

Weather forecast for users

Statistical post-processing (e.g. using KALMAN filters)

SWF Examples



Step 1

MicroArray Analysis

microarray data

Step 2

Clusfavor Analysis

Gene ID

Step 3

GenBank sequence retrieval

cDNA sequence

Step 4

NCBI BLAST search

genomic sequence

Step 5

Transfac search

Transcription factor binding

Step 6

Promoter Identification

promoter data

Step 7

Promoter Model generator

Consensus sequence

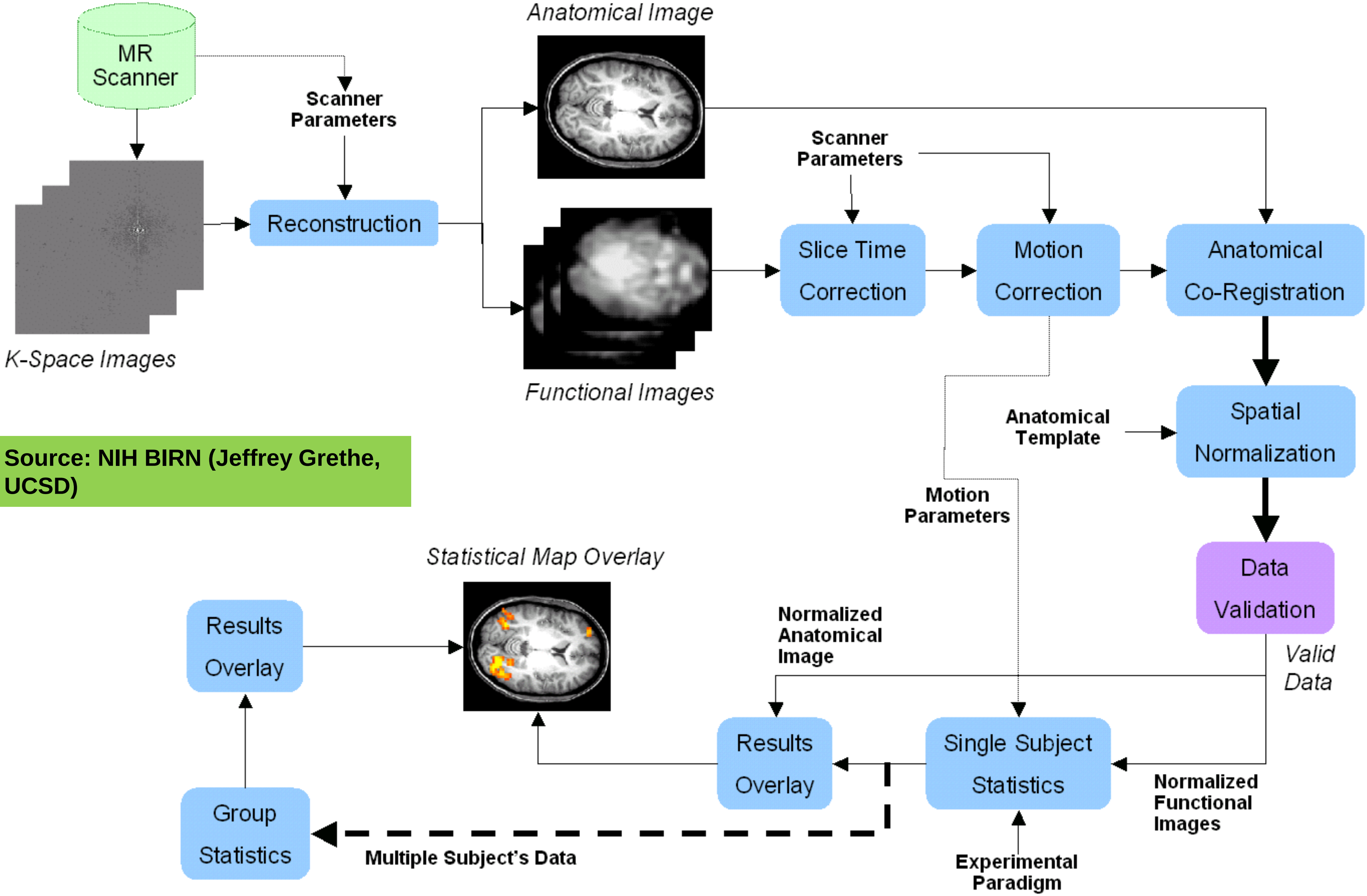
Step 8

NCBI BLAST search

new candidate target genes

Source: Matt Coleman (LLNL)

Functional MRI Analysis Workflow



MR Scanner

Scanner Parameters

K-Space Images

Reconstruction

Anatomical Image

Functional Images

Scanner Parameters

Slice Time Correction

Motion Correction

Anatomical Co-Registration

Anatomical Template

Spatial Normalization

Data Validation

Valid Data

Motion Parameters

Normalized Anatomical Image

Normalized Functional Images

Single Subject Statistics

Experimental Paradigm

Results Overlay

Statistical Map Overlay

Multiple Subject's Data

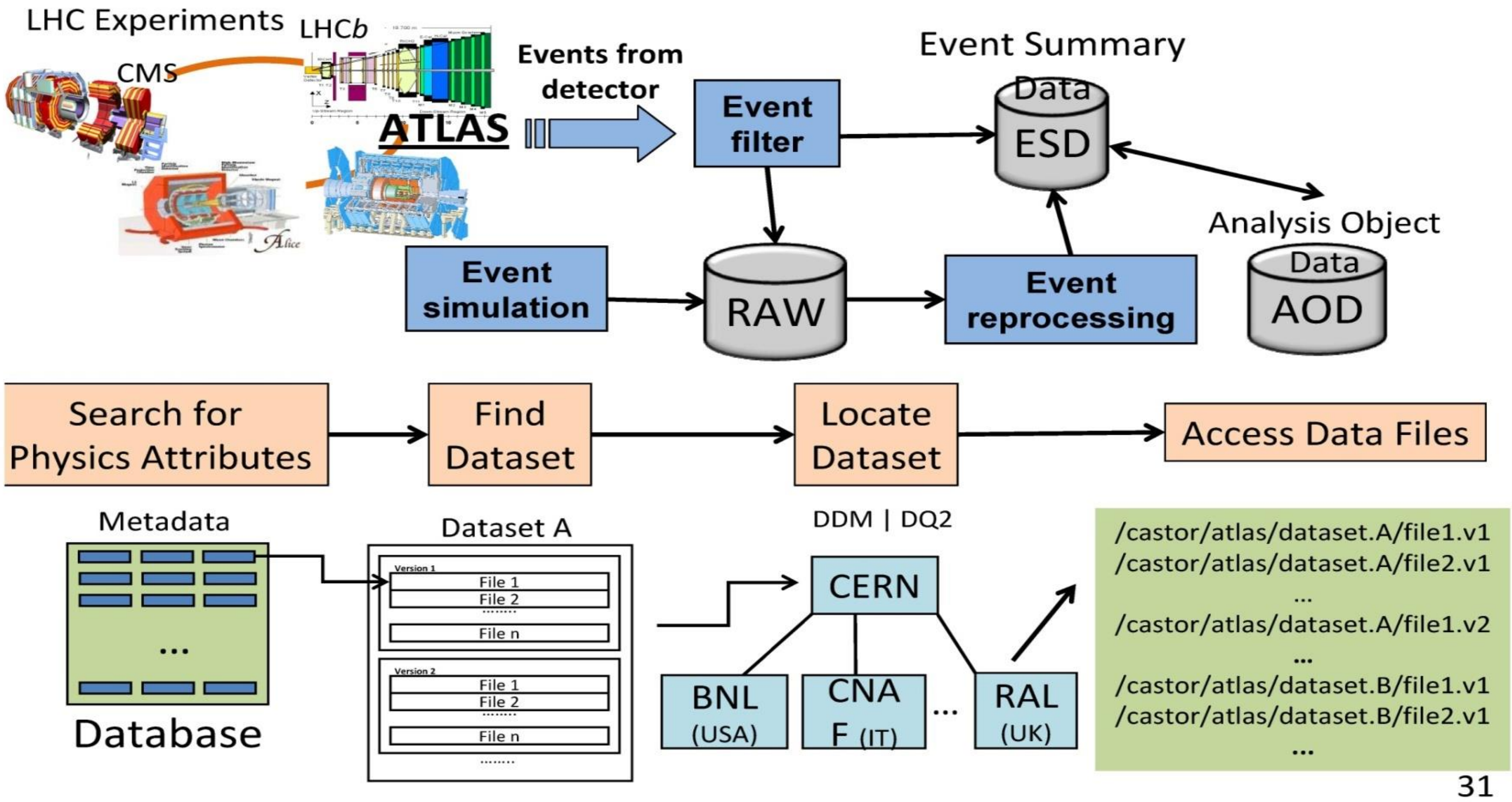
Group Statistics

Results Overlay

Source: NIH BIRN (Jeffrey Grethe, UCSD)

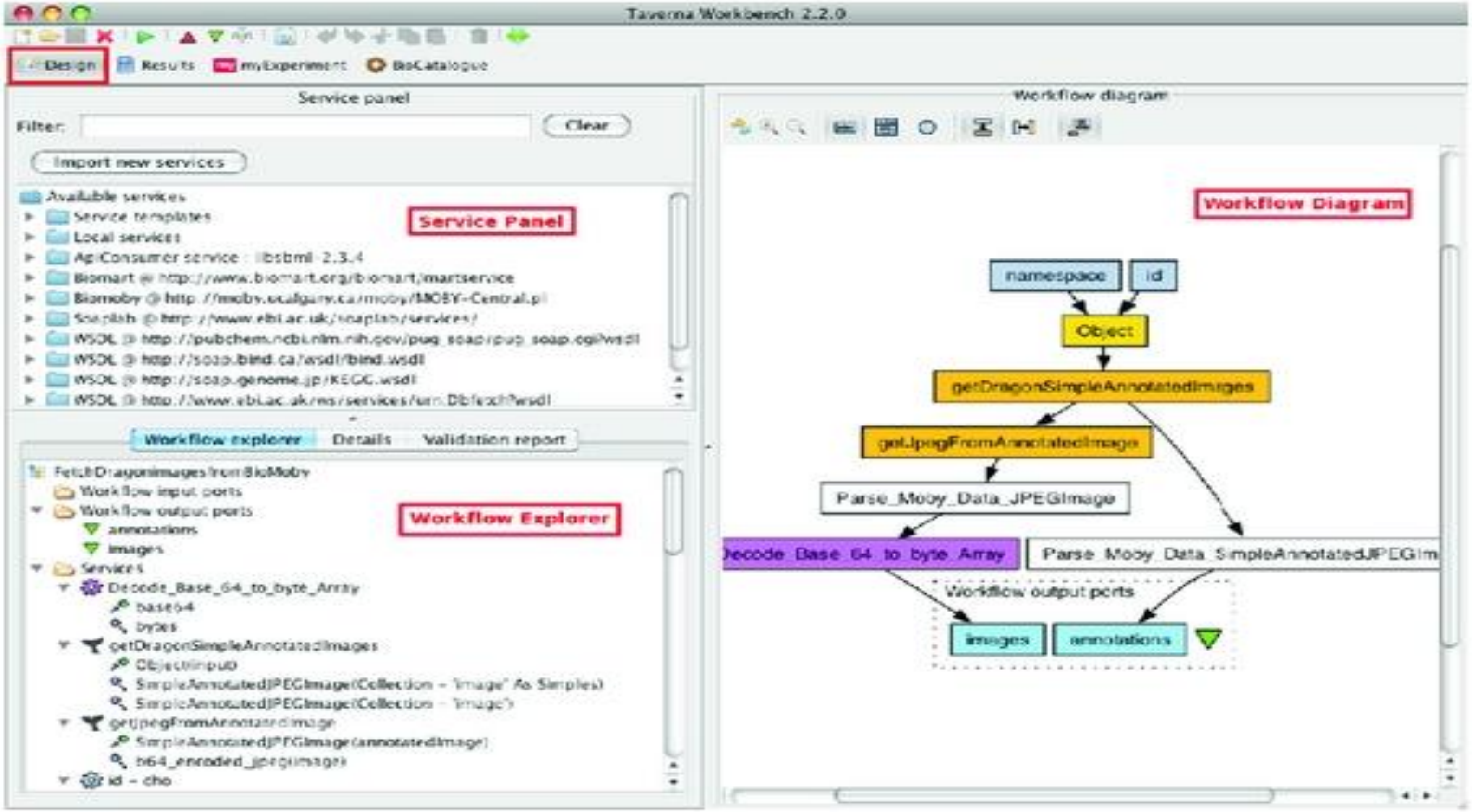
SWF Examples (cont)

ATLAS experiment (simplified)



/castor/atlas/dataset.A/file1.v1
/castor/atlas/dataset.A/file2.v1
…
/castor/atlas/dataset.A/file1.v2
…
/castor/atlas/dataset.B/file1.v1
/castor/atlas/dataset.B/file2.v1
…

SWF Examples (cont)


Taverna Workbench 2.2.0
Design
Results
myExperiment
BioCatalogue
Service panel
Filter:
Clear
Import new services
Available services
Service templates
Local services
ApiConsumer service : libsbml-2.3.4
Biomart @ http://www.biomart.org/biomart/martservice
Biomoby @ http://moby.ucalgary.ca/moby/MOBY-Central.pl
Soaplab @ http://www.ebi.ac.uk/soaplab/services/
WSDL @ http://pubchem.ncbi.nlm.nih.gov/pug_soap/pug_soap.cgi?wsdl
WSDL @ http://soap.bind.ca/wsdl/bind.wsdl
WSDL @ http://soap.genome.jp/KEGG.wsdl
WSDL @ http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl
Service Panel
Workflow explorer
Details
Validation report
FetchDragonimagesfromBioMoby
Workflow input ports
Workflow output ports
annotations
images
Services
Decode_Base_64_to_byte_Array
base64
bytes
getDragonSimpleAnnotatedImages
Object(input)
SimpleAnnotatedJPEGImage(Collection - 'Image' As Simples)
SimpleAnnotatedJPEGImage(Collection - 'Image')
getJpegFromAnnotatedImage
SimpleAnnotatedJPEGImage(annotatedImage)
b64_encoded_jpeg(image)
id - cho
Workflow Explorer
Workflow diagram
Workflow Diagram
namespace
id
Object
getDragonSimpleAnnotatedImages
getJpegFromAnnotatedImage
Parse_Moby_Data_JPEGImage
Decode_Base_64_to_byte_Array
Parse_Moby_Data_SimpleAnnotatedJPEGIm
Workflow output ports
images
annotations

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

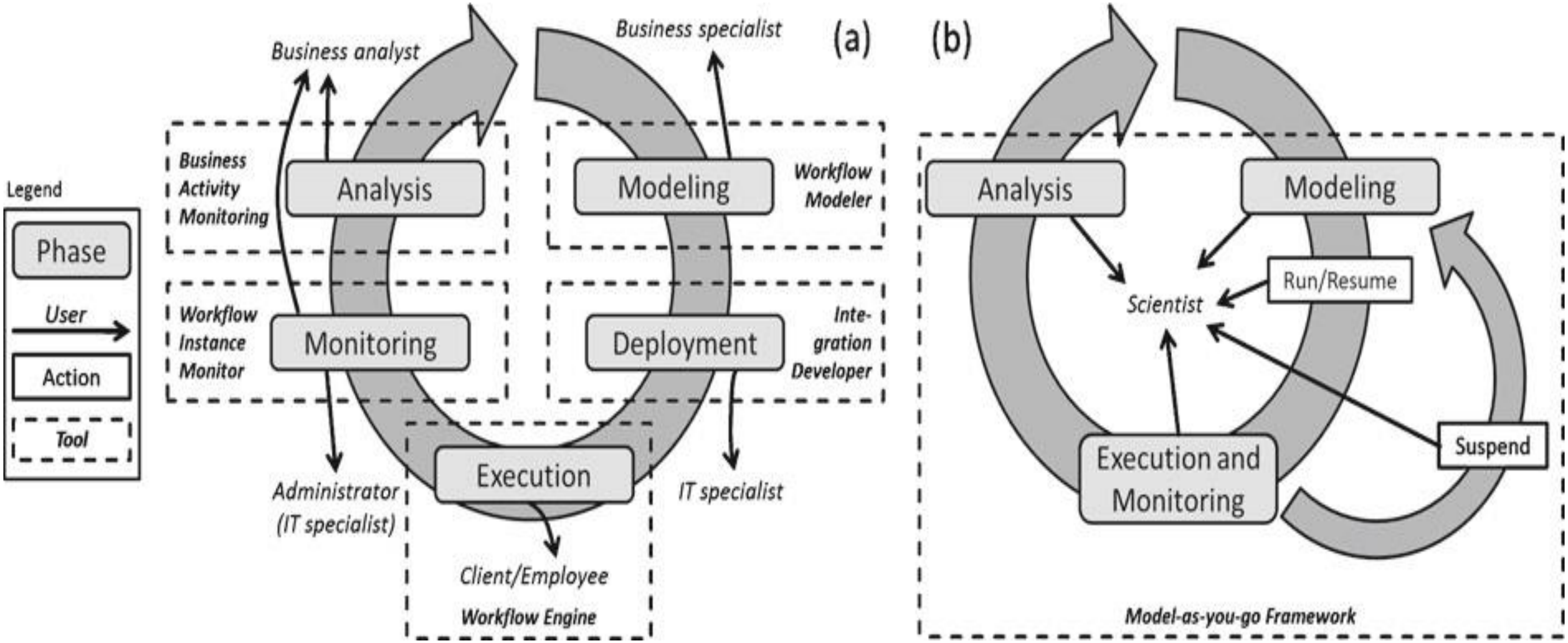
Scientific workflows vs. Business workflows

- 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

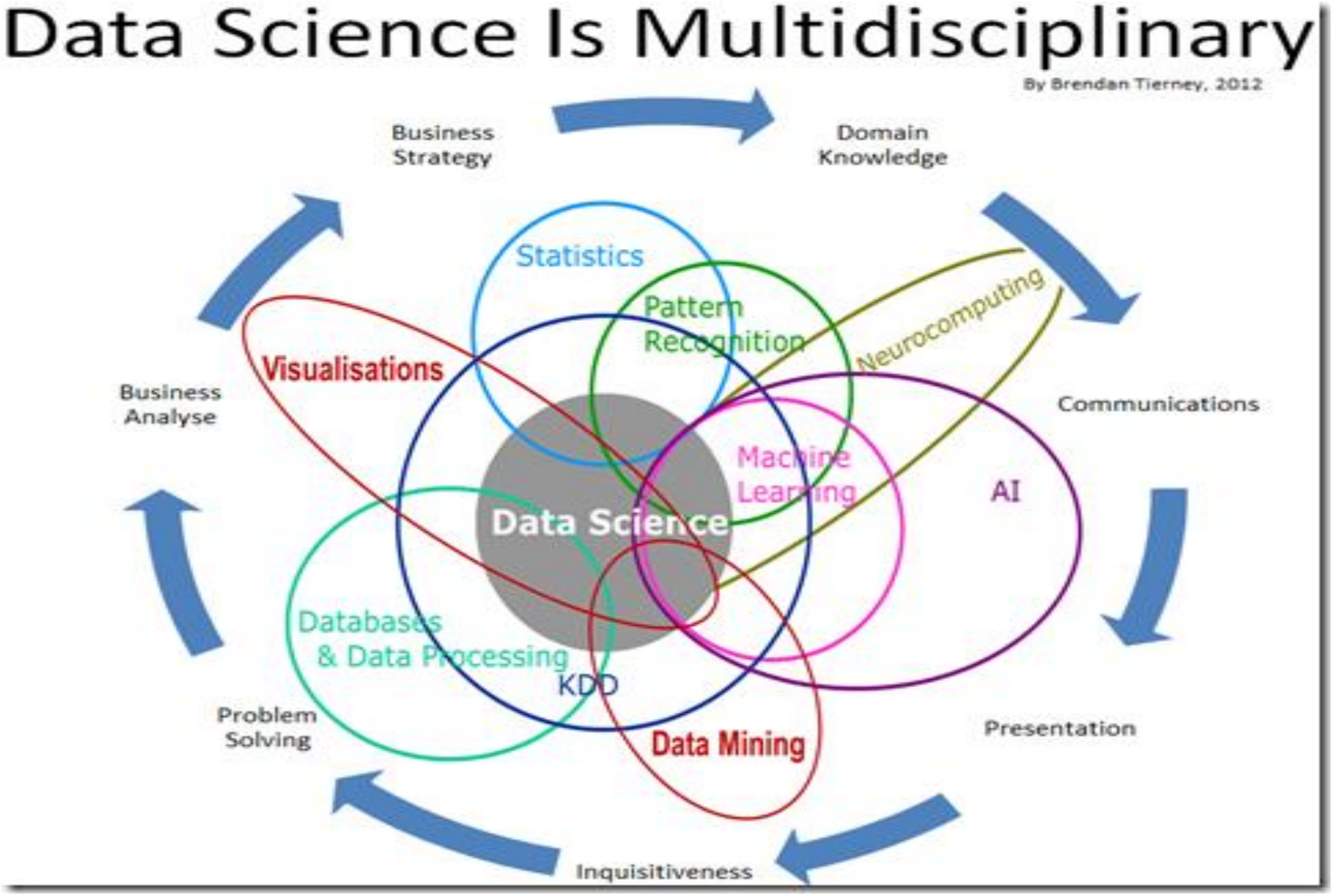
Scientific workflow and the scalability

- 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,

SWF: Programmable and reproducible Scalability

- Access and query data
- Scale computational analysis
- Increase reuse
- Save time, energy and money
- Formalize and standardize

Workflow, data science and big data



Workflow, data science and big data (cont)

- many SWfMSs are not prepared to handle large data sets because of inadequate support for distributed computing.
- most SWfMSs that do support distributed computing only allow static task execution orders

Workflow, data science and big data (cont)

- Develop new big data science technologies and infrastructure
- Develop data science workflow application through combination of tools, technologies and best practices
- Hands on consulting on workflow technologies for big data and cloud systems, e.g., MapReduce, Hdoop, Spark, Yann, Cascading.Oosie, Nova .
- Technology briefings and applied classes on end-to-end support for data science.

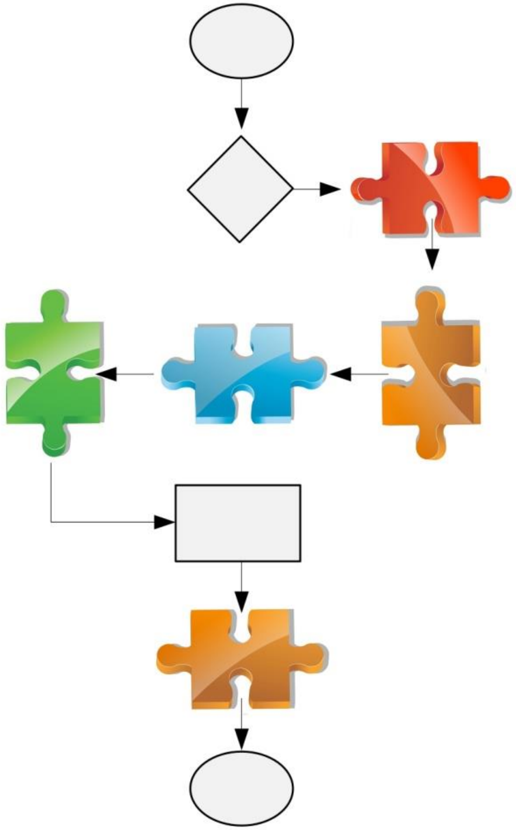
Some challenges

- Optimized execution on heterogeneous platforms.
- Representing and reasoning data: semantic, quality (unconsistent, uncertain)
- Increasing reuse within and across application domains
- Querying and integration of workflow provenance data

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)

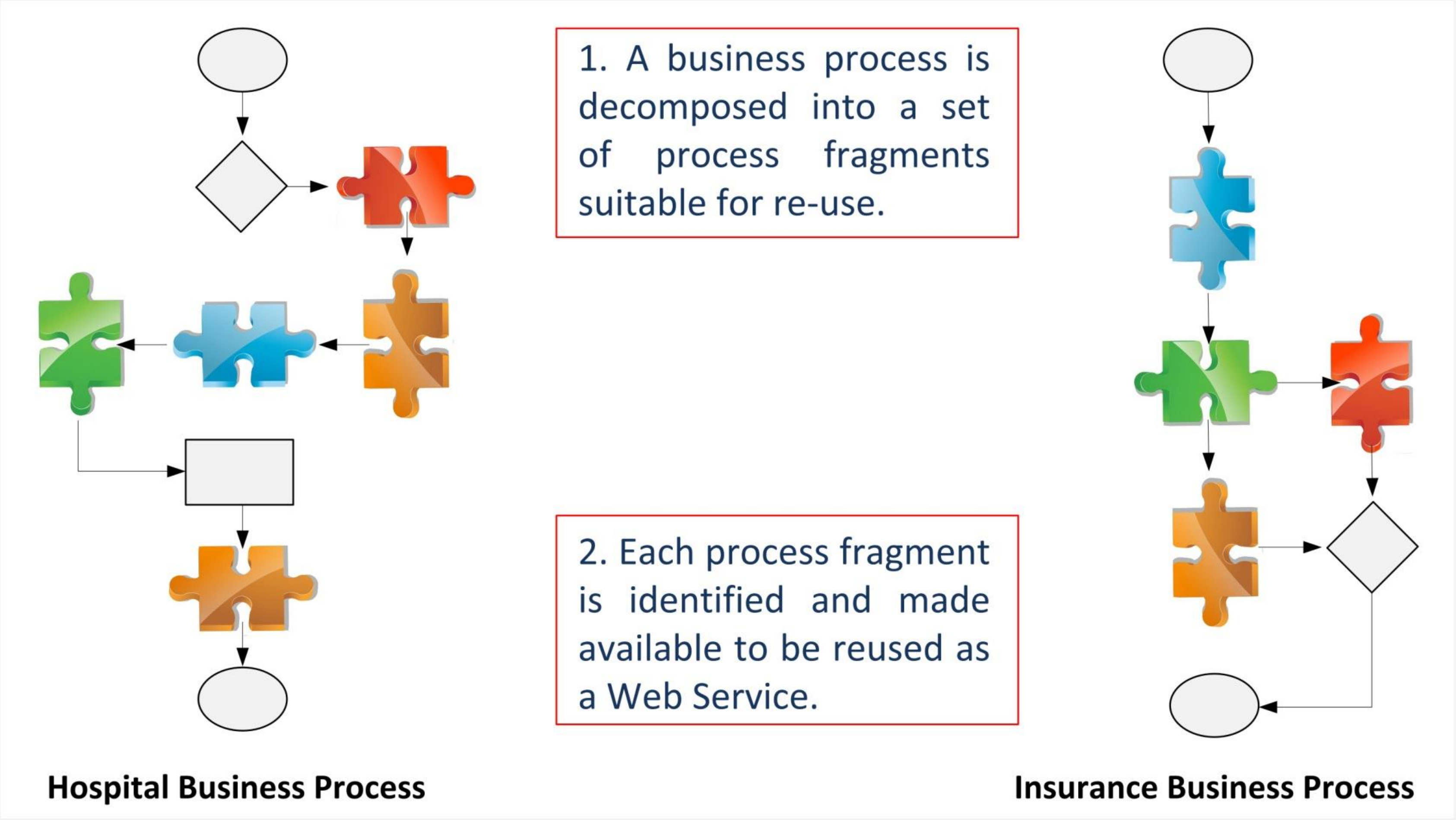
Decomposition of business process vs scientific process



Hospital Business Process

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

Identification of fragments



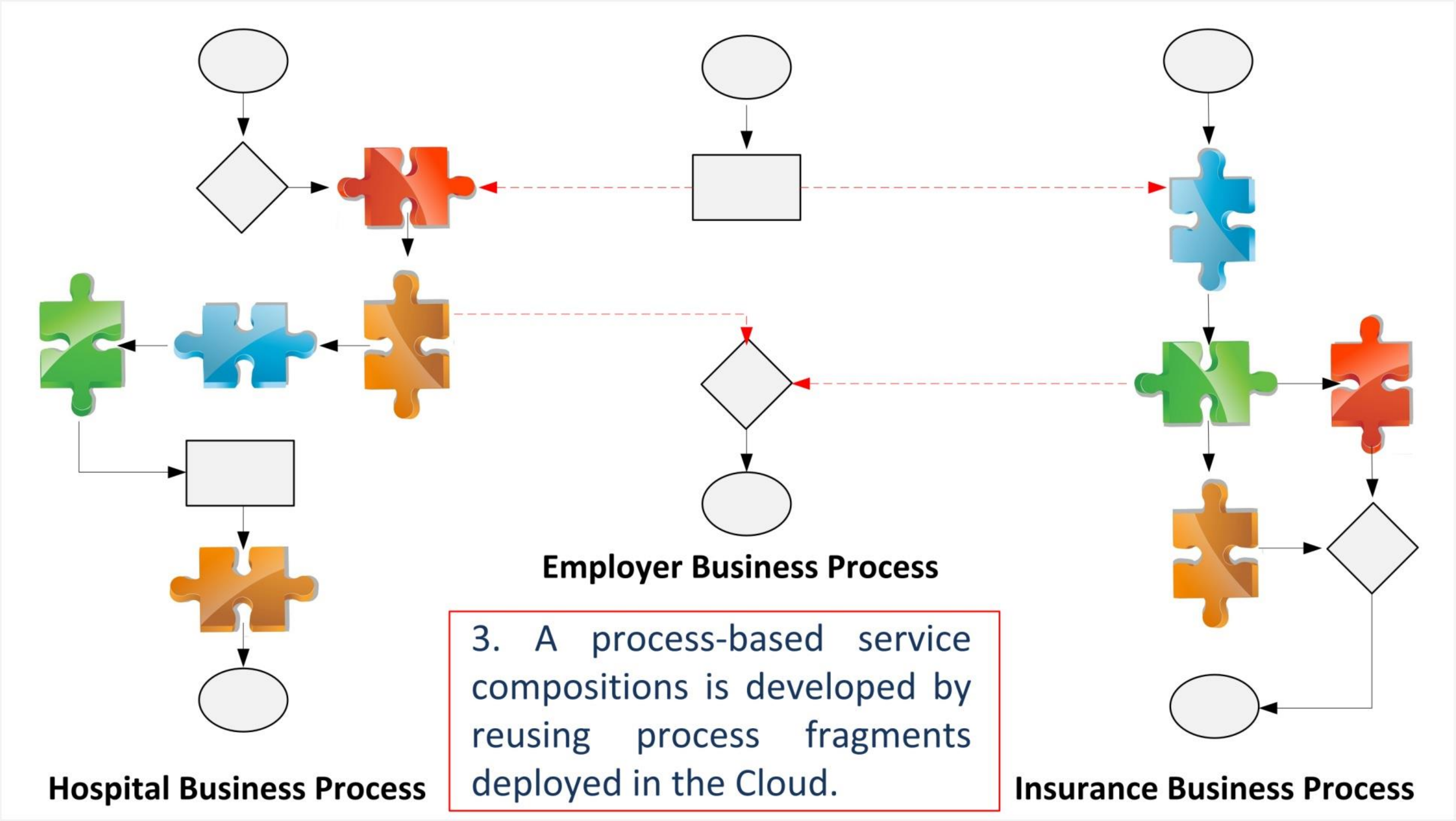
1. A business process is decomposed into a set of process fragments suitable for re-use.

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



Employer Business Process

3. A process-based service compositions is developed by reusing process fragments deployed in the Cloud.

Hospital Business Process

Insurance Business Process

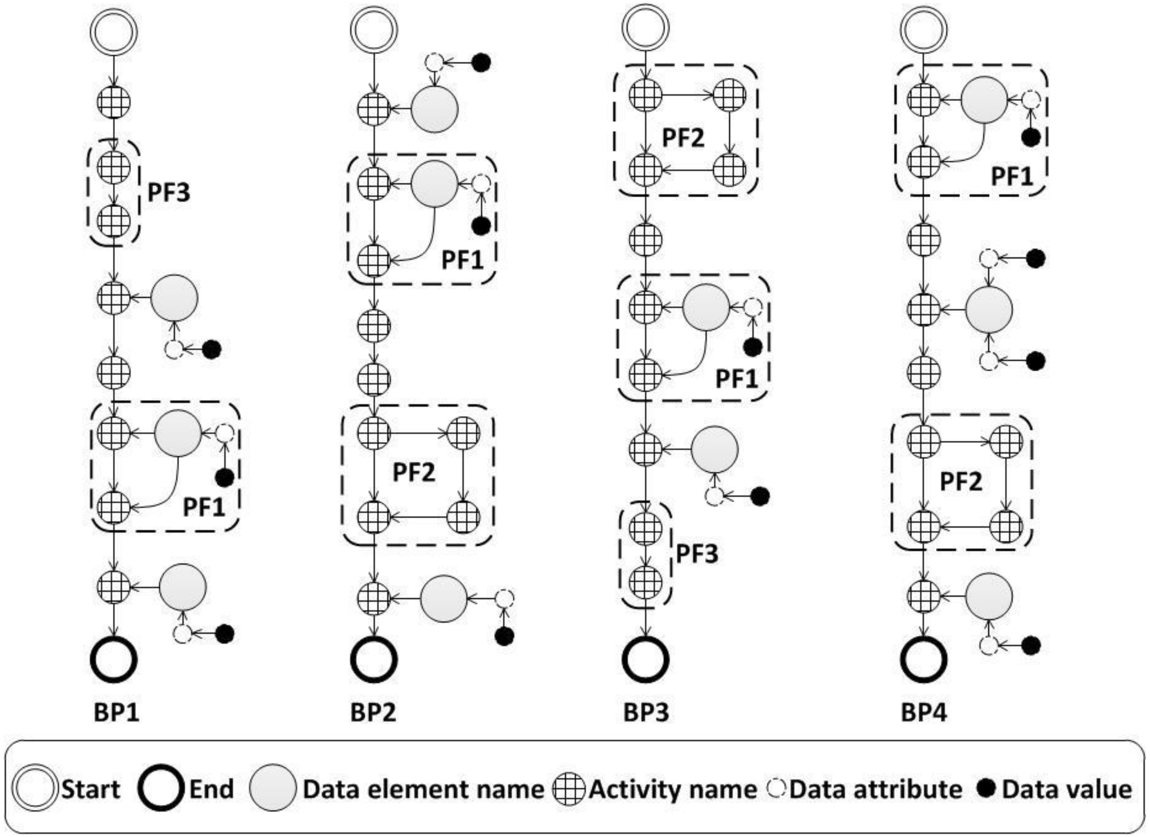
Provenance and privacy in SPaaS

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

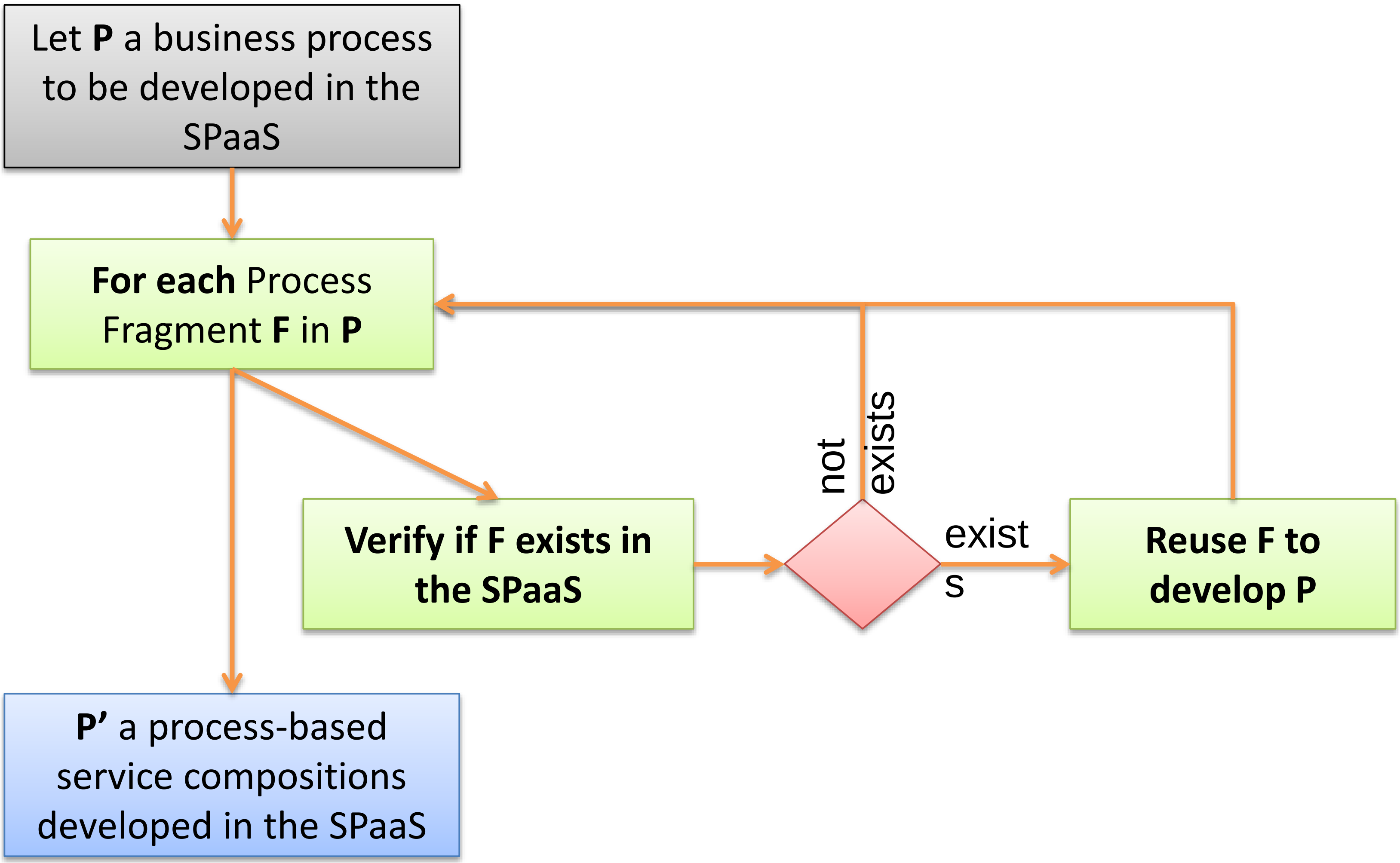
- **View on SPaaS:**

  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

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/