SimiFlow: Uma Arquitetura para Agrupamento de Workflows por Similaridade

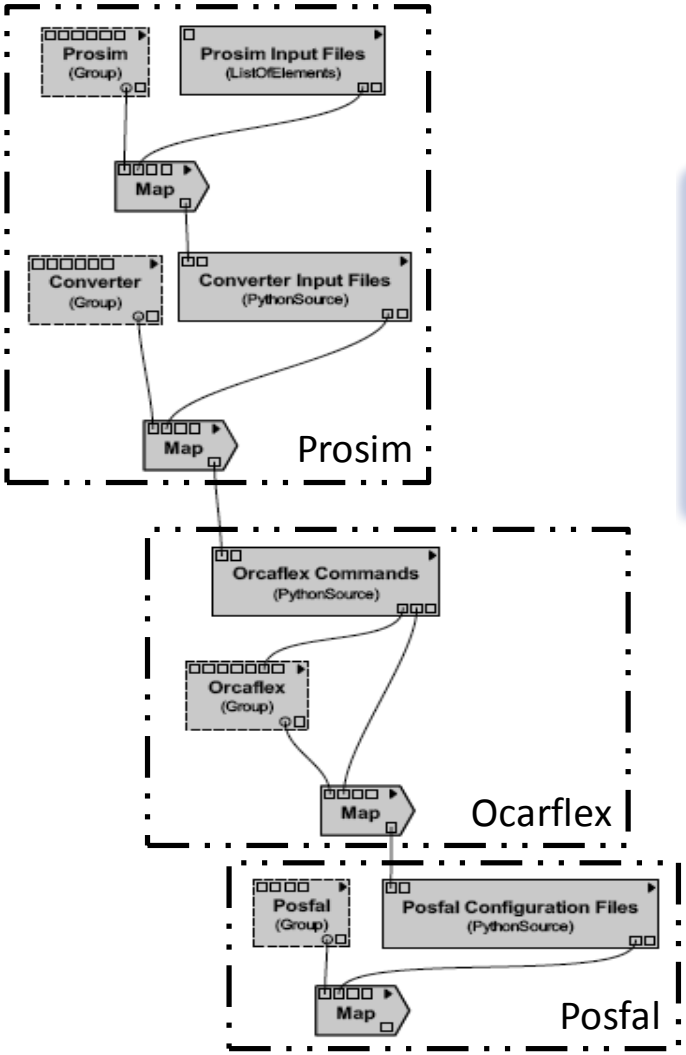
Vítor Silva, Fernando Chirigati, Kely Maia, Eduardo Ogasawara, Daniel de Oliveira, Vanessa Braganholo, Leonardo Murta, Marta Mattoso

IV e-Science Workshop – CSBC

Agenda

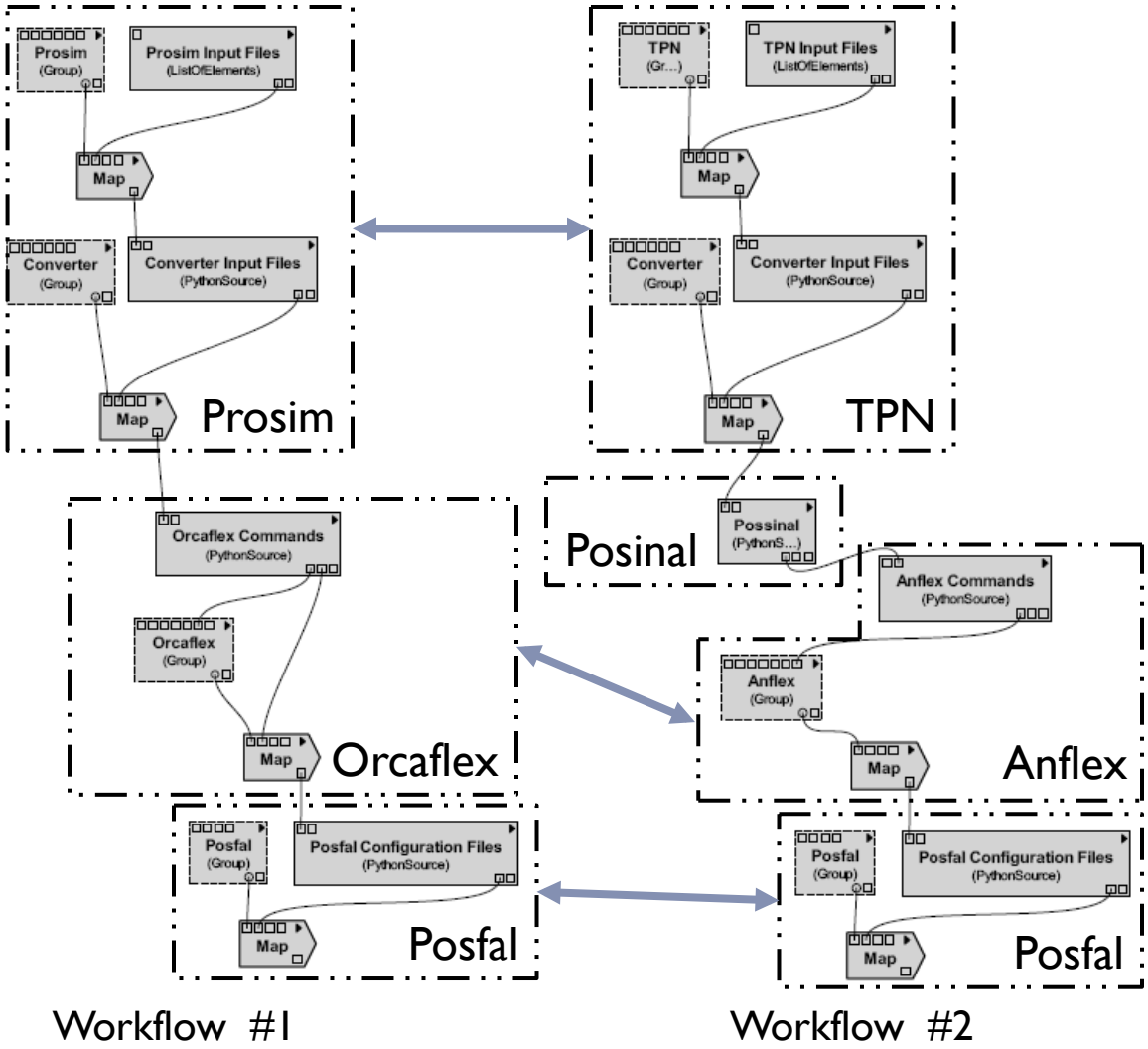
- Workflow Científico
- Workflow Conceitual x Workflow Concreto
- Linhas de Experimentos
- SimiFlow
- Resultados Preliminares
- Trabalhos Relacionados
- Conclusão
- Trabalhos Futuros

Workflow Científico



Um workflow científico é caracterizado por uma coleção de atividades organizada de modo a representar um modelo científico.

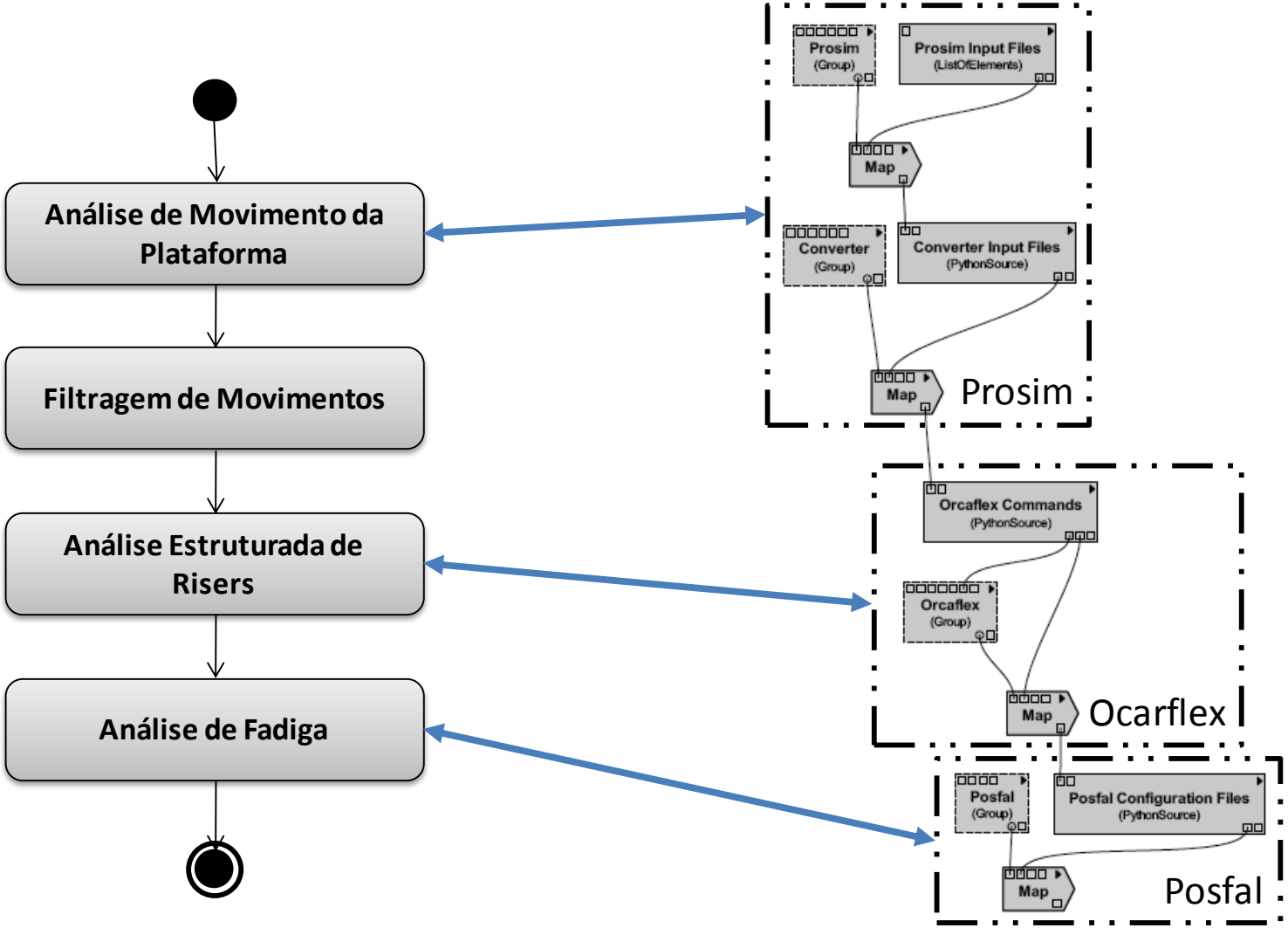
Workflow Concreto para Exploração de Petróleo em Águas Profundas



Workflow #1

Workflow #2

Workflow Conceitual x Workflow Concreto



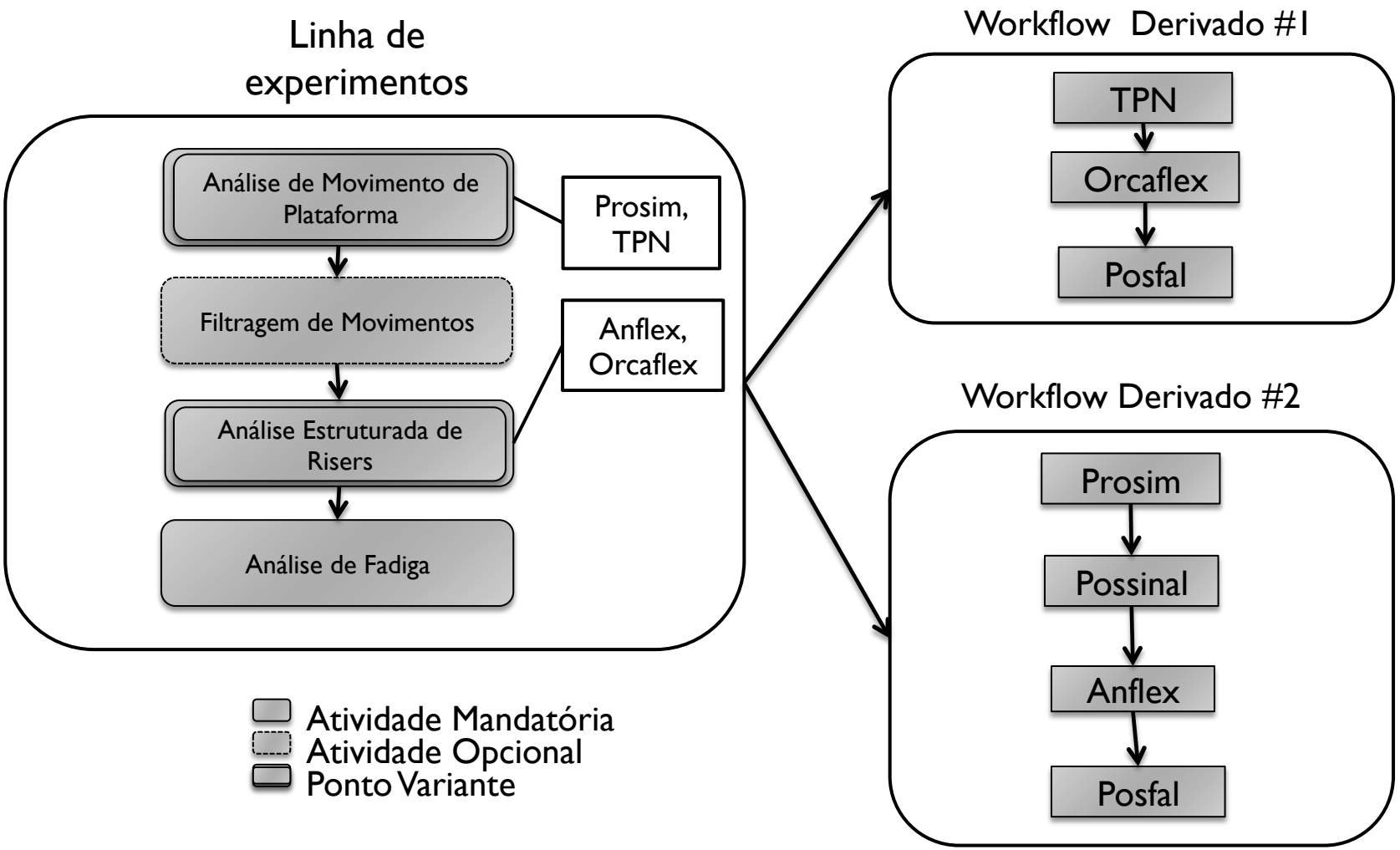
Os cientistas precisam de apoio...

- Para uma representação independente que
  - Apresenta a abstração do workflow baseado em conceitos
  - Apresenta alternativas para o experimento
  - Apresenta as atividades que não podem ser executadas no experimento
- Para capturar a proveniência prospectiva, considerando-se os conceitos do experimento científico
  - Agrupamento de diferentes workflows científicos como parte de um único experimento para futuras análises

Linhas de Experimentos

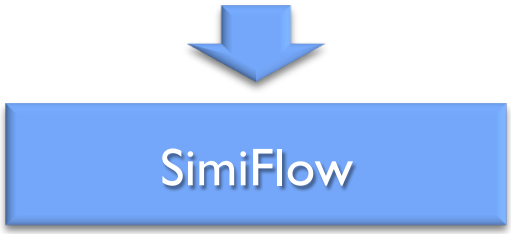
- Apoio à composição de workflows científicos
- Definição de famílias de experimentos
  - Linhas de experimentos atuam com abstrações utilizando terminologia do domínio representado
  - Representam todas as possibilidades e variações dos experimentos de um determinado domínio
  - Pode-se derivar workflows concretos a partir de uma linha de experimentos
- Ferramenta GExpLine
  - Responsável pela gerência de linhas de experimentos

Linhas de Experimentos



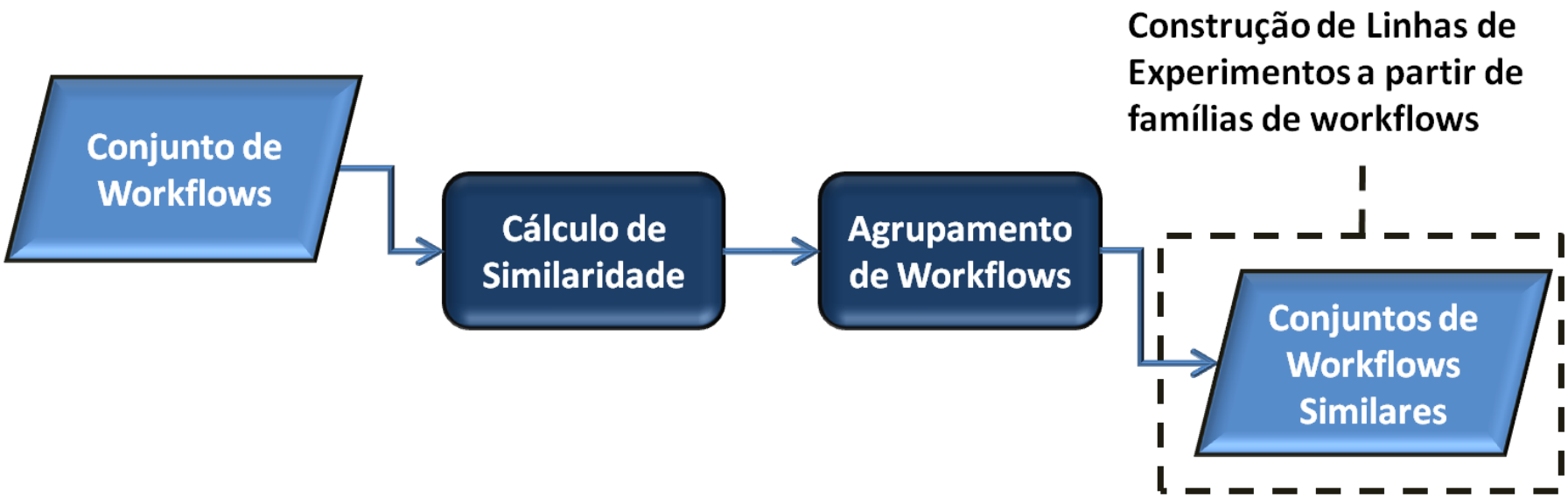
Linhas de Experimentos

- Criação de linhas de experimentos
  - Descendente (conceitual → concreto)
    - Cientistas iniciam a concepção de um novo experimento
    - Não há workflows concretos disponíveis
  - Ascendente (concreto → conceitual)
    - Presença de workflows pré-existentes
    - Diversos workflows concretos são representações de um mesmo workflow conceitual
    - Exige comparação entre os workflows concretos, cálculo da similaridade e agrupamento para criar famílias de experimentos



SimiFlow - Arquitetura

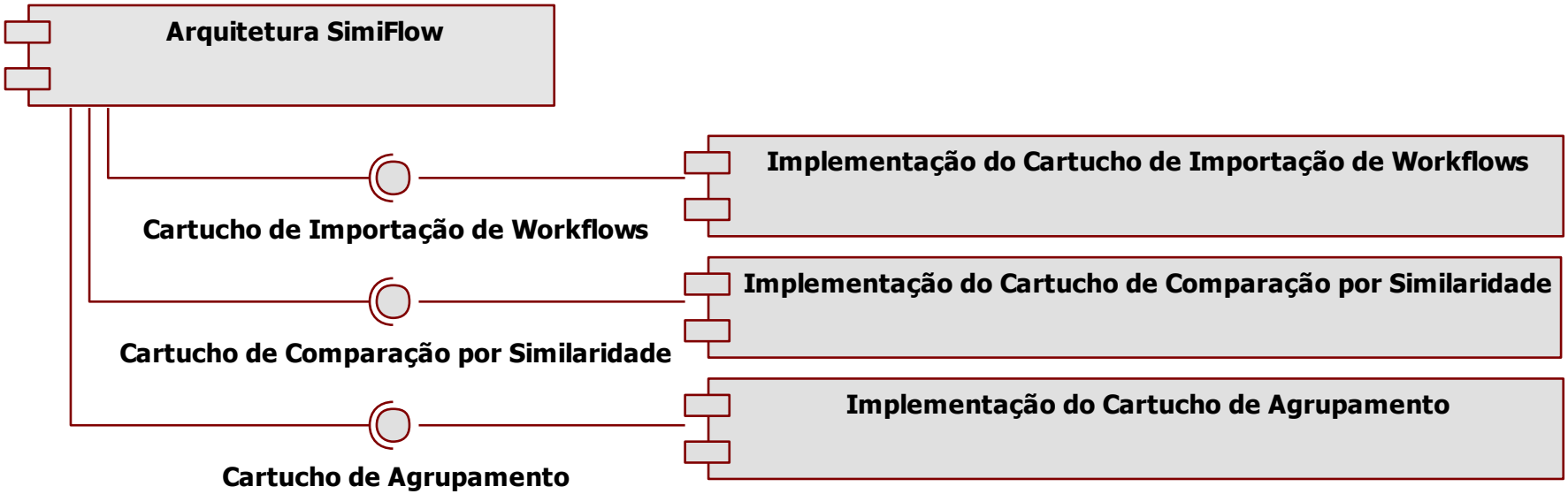
- A partir de um conjunto de workflows, é realizado o cálculo da Similaridade para todos os pares de elementos deste conjunto
- É gerado um Agrupamento de workflows através da Similaridade calculada



SimiFlow - Arquitetura

- Divisão em três cartuchos
  - Importação de Workflows
  - Comparação por Similaridade
  - Agrupamento

Arquitetura baseada no Design Pattern Strategy.



SimiFlow – Importação de Workflows

- Apóia o armazenamento das propriedades de workflows provenientes dos seguintes SGWfC
  - Taverna
  - VisTrails
  - Kepler
- Uso do sítio myExperiment
  - Importação de todos os workflows da base de dados do sítio
    - SGWfC Taverna
    - 459 workflows públicos

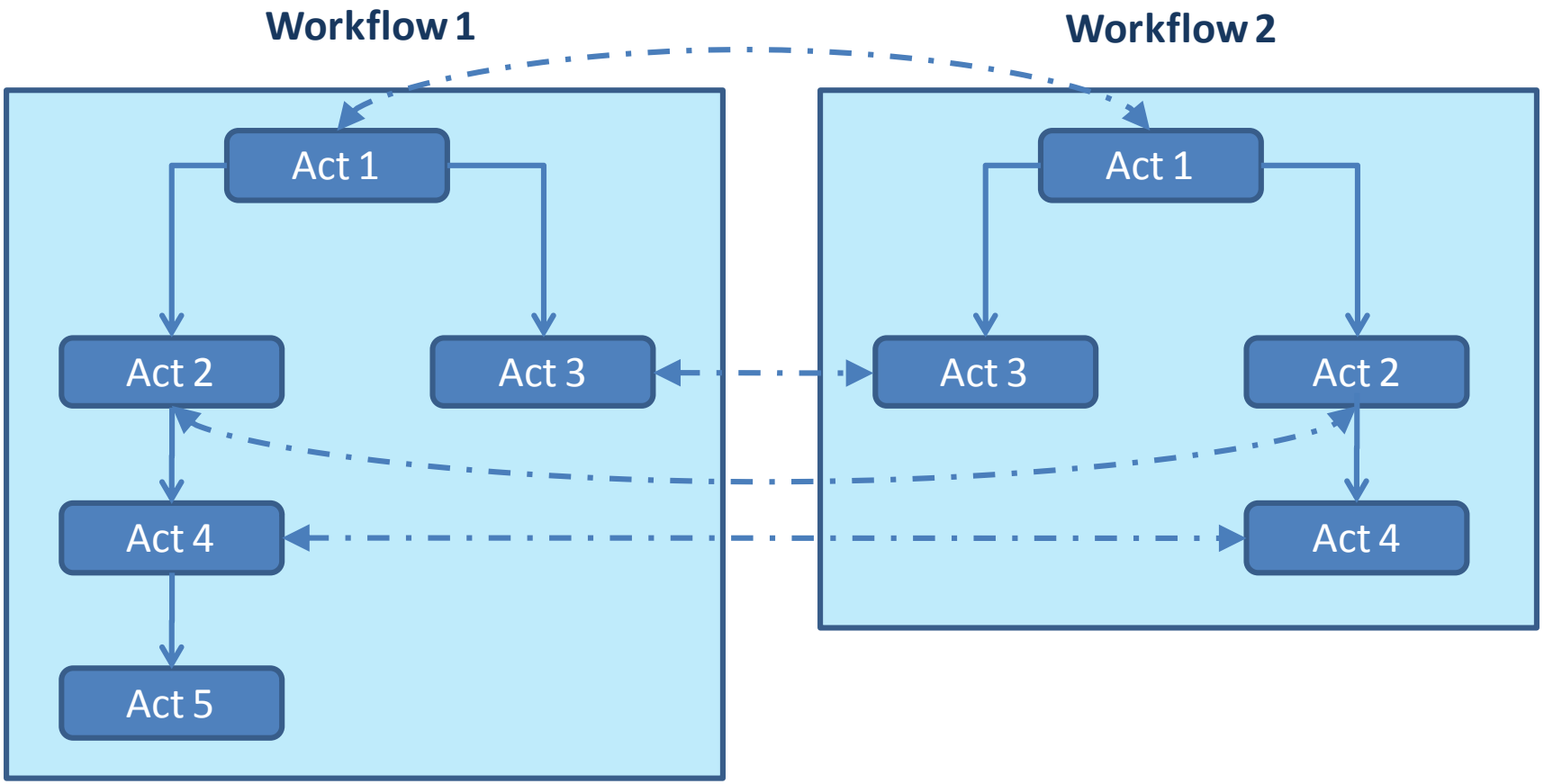
SimiFlow – Comparação por Similaridade

- Semelhança entre workflows
  - Combina pares de atividades entre dois workflows
  - Considera as propriedades das atividades em um casamento de atividades
    - Nome
    - Portas de entrada e de saída
    - Relacionamentos
    - Existência de subworkflows

SimiFlow – Comparação por Similaridade

- Cálculo da Similaridade
  - Baseada em um sistema de penalização
  - Para cada par de atividades, há uma análise de suas propriedades específicas
    - Cada propriedade diferente estabelece uma penalização no valor da Similaridade
  - O valor da penalização está relacionada à importância da propriedade

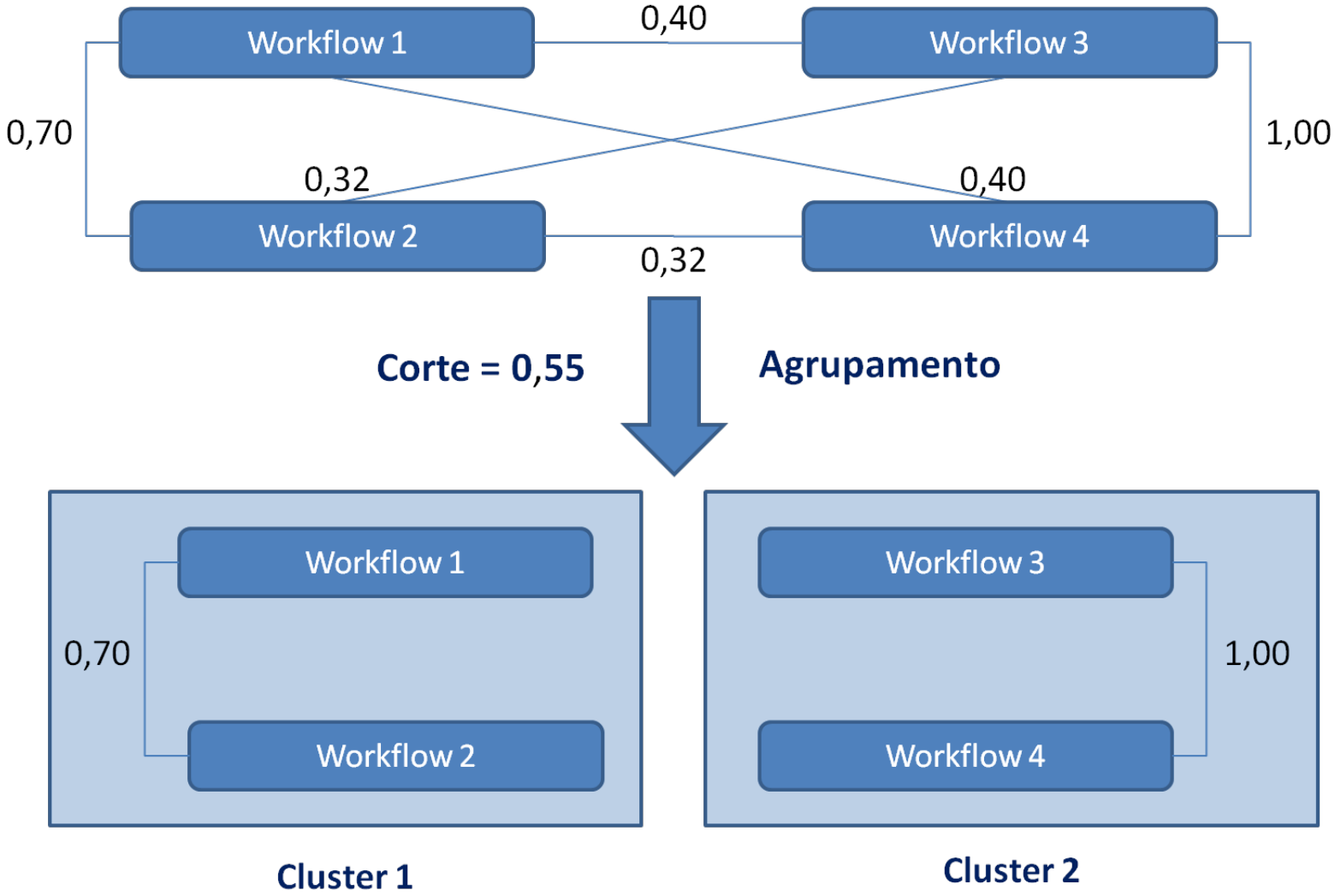
Comparação por Similaridade - Exemplo



SimiFlow – Agrupamento

- Utiliza o cálculo de similaridade de todos os pares de workflows presentes na base de dados
- Cria um grafo completo a partir dos valores de similaridade (arestas) e dos workflows (vértices)
- Realiza o corte das arestas com um valor de similaridade inferior a um valor de corte
- Cria componentes conexos – subgrafos conexos em uma floresta

SimiFlow - Agrupamento



SimiFlow – Estudos Preliminares

- Dois estudos foram realizados
- Primeiro estudo
  - Preocupação em avaliar a capacidade de realizar os agrupamentos
  - Uso de 24 workflows do Taverna (seleção manual)
  - Resultados satisfatórios

Nome do *Workflow*	Grupo	Nome do *Workflow*	Grupo
Discover_entities	1	Get Concepts	4
Retrieve_documents_MR1	1	Get CVs	4
EBI_ClustalW2	2	Get Evidence Types	4
EBI_ClustalW2_phylogentic_tree	2	Get graphs	4
EBI_Kalign	3	Get Graphs of Name	4
EBI_TCoffee	3	Get relations	4

SimiFlow – Estudos Preliminares

- Segundo estudo
  - Preocupação em avaliar a escalabilidade da arquitetura
  - Geração de 459 workflows pela importação
    - Total de 105.111 comparações (C_2^{459}
  - Resultados
    - Agrupamento precisa de ajustes
    - Calibragem das penalizações ainda precisa ser refinada
    - Necessidade do uso da distribuição e do paralelismo para realizar as comparações
      - Cartucho de Comparação por Similaridade é pesado computacionalmente

Trabalhos Relacionados

- Santos *et al.* (2008)
  - Agrupamento de workflows
  - Representações por grafo e por vetor apresentam algumas limitações
    - Propriedades específicas dos workflows não são consideradas
- Ohst *et al.* (2003)
  - Diferenças entre versões de um documento UML
- Uhrig (2008)
  - Objetivo de igualar dois diagramas de classes
- SiDiff (2010)
  - Comparação de modelos de dados
  - Existência dos pesos é equivalente à penalização apresentada

Conclusão

- O SimiFlow pode facilitar a organização de bases extensas de workflows científicos pelos agrupamentos encontrados
- Apoio na construção de linhas de experimentos por workflows que pertençam a mesma família (ascendente)
- A partir dos estudos preliminares
  - Foi possível identificar agrupamentos de workflows semelhantes pelo primeiro estudo, apesar dos pontos a serem melhorados
  - No segundo estudo, verificou-se a necessidade de calibragem das penalizações no processo de comparação por similaridade

Trabalhos Futuros

- Depuração dos métodos já implementados
- Testes com outros algoritmos de comparação por similaridade e de agrupamento
  - Construção de novos cartuchos
- Uso do paralelismo para melhorar o desempenho do processo de comparação

SimiFlow: Uma Arquitetura para Agrupamento de Workflows por Similaridade

Obrigado!

Vítor Silva, Fernando Chirigati, Kely Maia, Eduardo Ogasawara, Daniel de Oliveira, Vanessa Braganholo, Leonardo Murta, Marta Mattoso

IV e-Science Workshop – CSBC