Gerência de Experimentos Científicos em Nuvens de Computadores: Oportunidades de Pesquisa em Banco de Dados

Daniel de Oliveira – danielcmo@ic.uff.br

Marta Mattoso – marta@cos.ufrj.br

Quem nós somos

- Professora associada 4 no Programa de Engenharia de Sistemas e Computação da COPPE/UFRJ
- Bolsista de produtividade em pesquisa do CNPq nível 1C.
- Publicou mais de 200 artigos completos em revistas e congressos.
- Foi diretora de publicações da SBC no período de 2005 a 2007.
- Coordena diversos projetos de pesquisa com financiamento do CNPq, Capes, INRIA e FAPERJ
- Atua principalmente nas áreas de distribuição e paralelismo em bancos de dados, workflows científicos em ambientes de paralelismo e gerência de dados de proveniência.

Profa. Marta Mattoso, D.Sc.

COPPE
UFRJ

Quem nós somos

- Professor adjunto do Instituto de Computação da Universidade Federal Fluminense (UFF) desde 2013.
- Recebeu o grau de Doutor em Ciências pela UFRJ em 2012.
- Publicou mais de 50 artigos em periódicos indexados e em congressos nacionais e internacionais.
- Seus interesses de pesquisa incluem bancos de dados, computação em nuvem, gerência de workflows científicos, paralelismo de dados, bioinformática e mineração de dados.
- É membro da ACM, IEEE e SBC.

Prof. Daniel de Oliveira, D.Sc.

Instituto de Computação

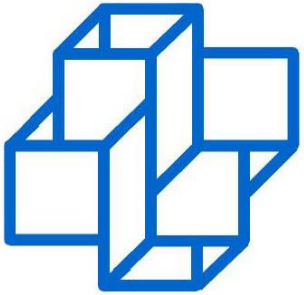
Grupo de Pesquisas

- **Marta Mattoso, D.Sc. – COPPE/UFRJ**
- **Daniel de Oliveira, D.Sc. – IC/UFF**
- Eduardo Ogasawara, D.Sc. – CEFET/RJ
- Pós-doutorandos
  - Kary Ocaña, D.Sc. – COPPE/UFRJ
- Alunos de D.Sc.
  - Jonas Dias
  - Flavio Costa

Colaborações Atuais

- INRIA
  - Patrick Valduriez

- LNCC
  - Fabio Porto
  - Luiz Gadelha Jr.



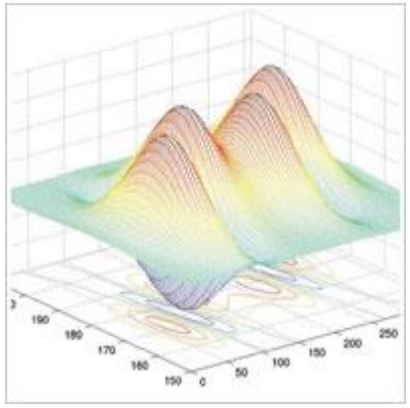
Laboratório Nacional de Computação Científica

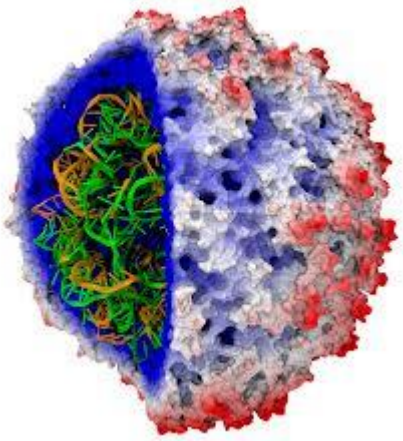
Roteiro do Tutorial

- **Motivação**
- Workflows Científicos
- Nuvens de Computadores
- Proveniência de Dados
- Máquinas de Workflow para Nuvem
- Aplicação de Proveniência em e-Science
- Demo

e-Science

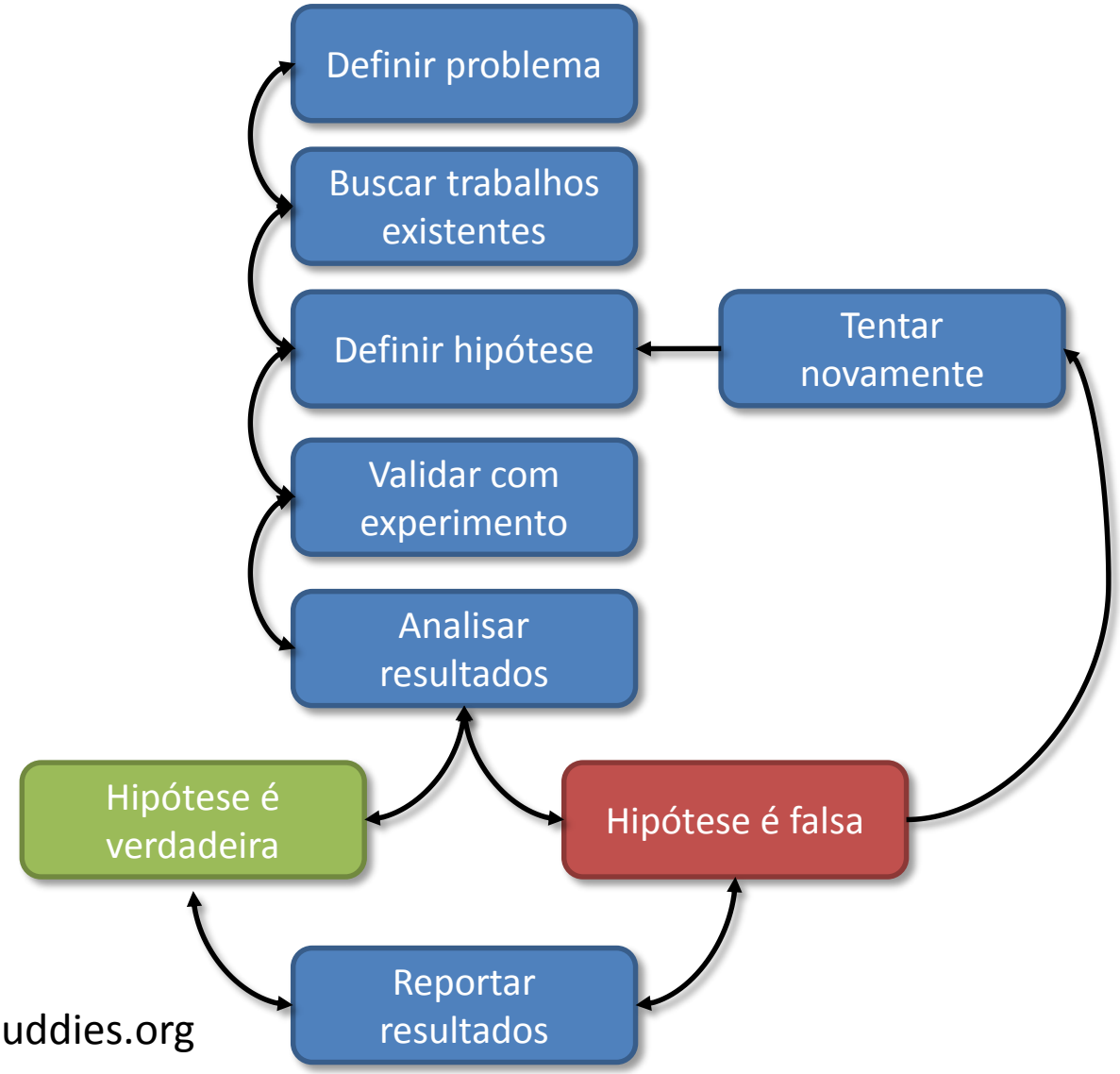
- "A combinação de pesquisa em computação e modelagem matemática que proporciona a aceleração da pesquisa em outras áreas do conhecimento"





Fonte: www2.bioetanol.org.br/escibioenergy

Método Científico



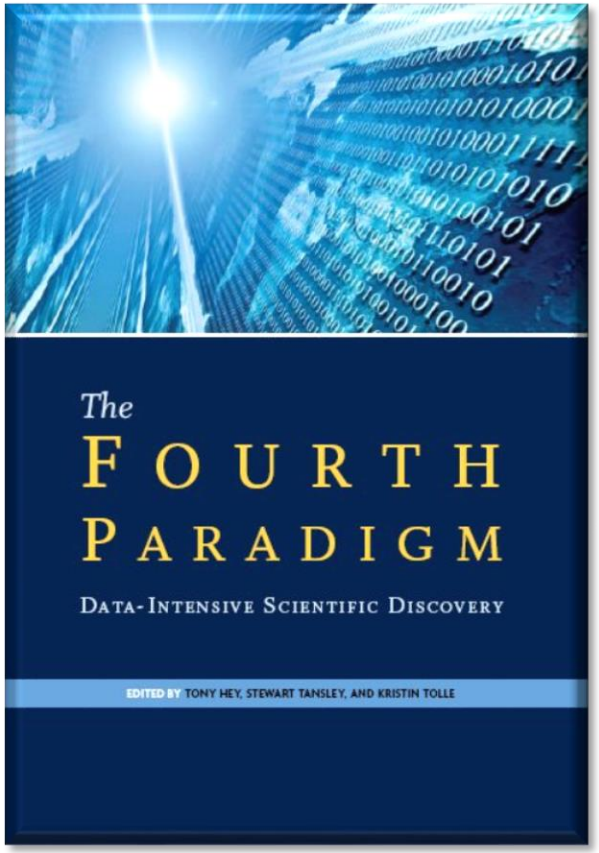
http://www.sciencebuddies.org

Método Científico *in silico*

- "Ten years ago, the problem was how to obtain data."
- Nos dias de hoje o gargalo se encontra em determinar quais estratégias computacionais serão utilizadas para que o cientista possa analisar esse grande volume de dados produzido e coletado
- Dados geralmente são heterogêneos e distribuídos

Fonte: www2.bioetanol.org.br/escibioenergy

O quarto paradigma



"The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of e-Science such as **databases**, workflow management, visualization, and **cloud computing** technologies."

http://research.microsoft.com/en-us/collaboration/fourthparadigm/

O quarto paradigma

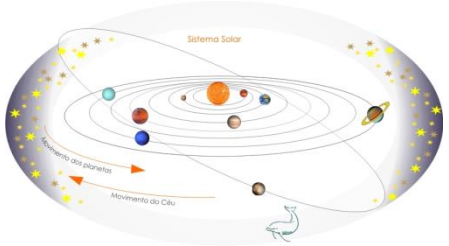
"The speed at which any given
ances
its
with
of e-
ses,
"

O dado não pode mais ser gerenciado separadamente do processo que o gerou!

http://research.microsoft.com/en-us/collaboration/fourthparadigm/

e-Science

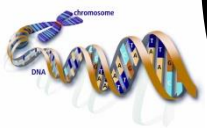
Medicina



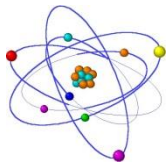
Astronomia

e-Science

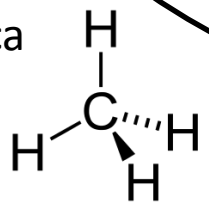
Biologia



Física



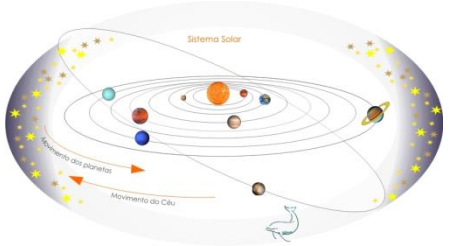
Química



Engenharias

e-Science

Medicina



Astronomia

Banco de Dados

Redes

Engenharia de Software

e-Science

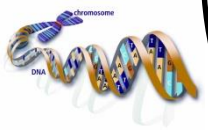
Segurança

HPC

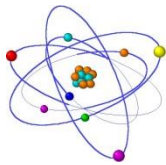
Algoritmos

Inteligência Artificial

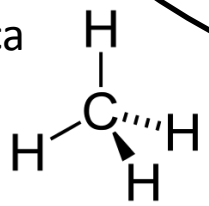
Biologia



Física



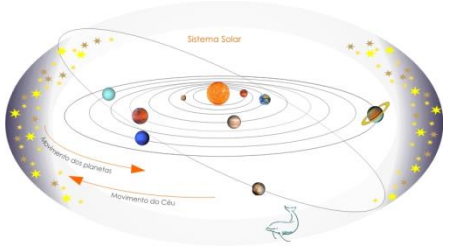
Química



Engenharias

e-Science

Medicina



Astronomia

Banco de Dados

Engenharia de Software

Redes

e-Science

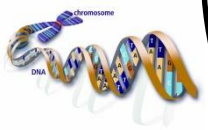
Segurança

HPC

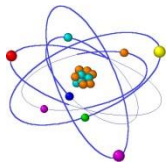
Algoritmos

Inteligência Artificial

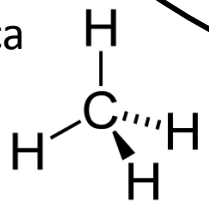
Biologia



Física



Química



Engenharias

e-Science no Reino Unido

- E-Science é vital para a exploração do poderoso aparato científico da próxima geração
  - Telescópios
  - Sensores
  - Satélites

- Cada um desses aparatos podem gerar vários terabytes de dados diariamente
- Os usuários podem estar dispersos geograficamente ao redor do globo
- A e-Science deve construir a infraestrutura necessária para este exploração

e-Science no Brasil

- São Paulo School of Advanced Science on e-Science for Bioenergy Research

- Grandes Desafios da Computação no Brasil
  - http://www.gd2.ufam.edu.br/
- BreSci - Brazilian e-Science workshop
  - http://www.ic.ufal.br/csbc2013/noticias/e-science

e-Science no Brasil

FAPESP
FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO

Página inicial » A Instituição » Eventos

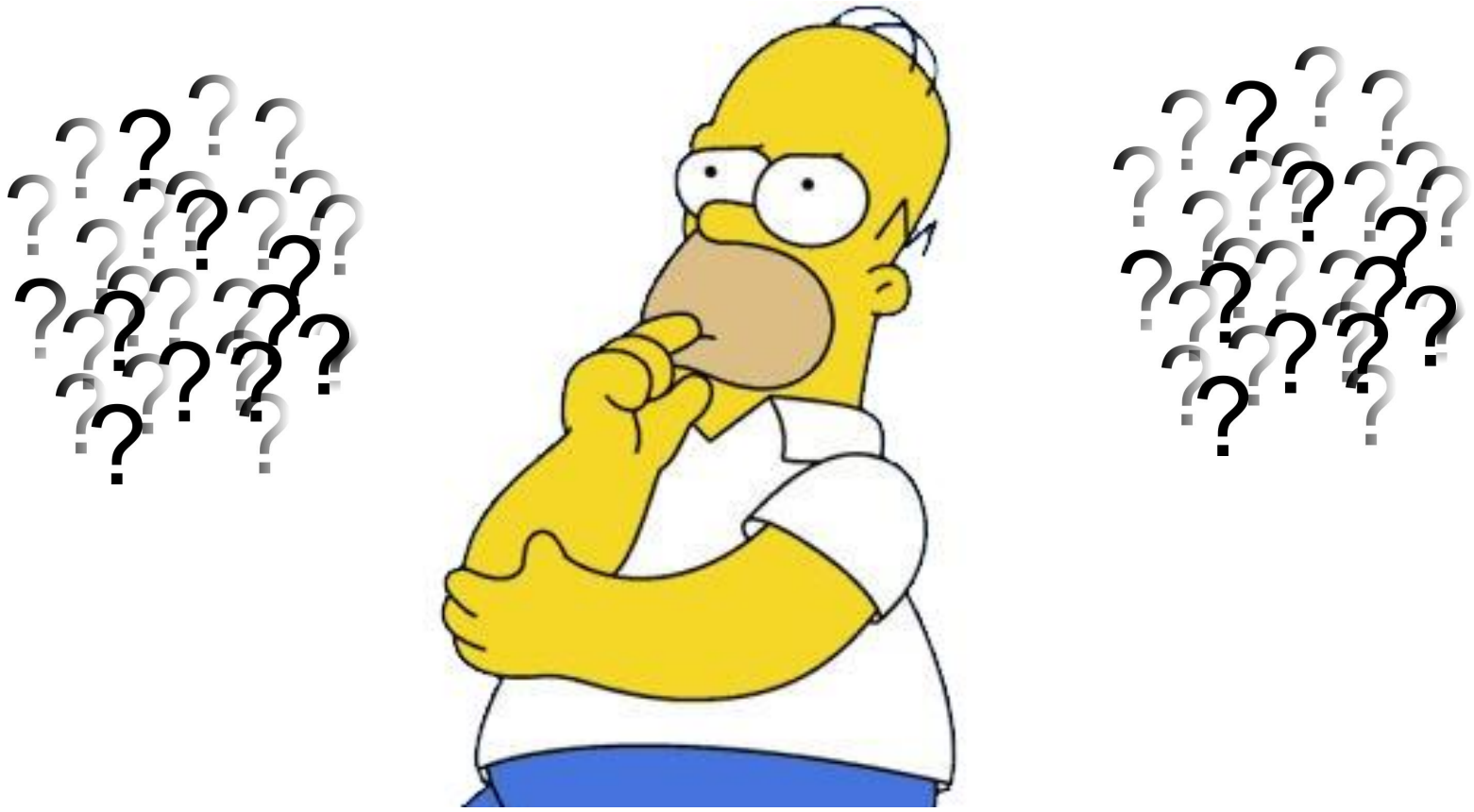
Latin American eScience Workshop 2013

Sponsored jointly by Microsoft Research and FAPESP

May 13–15, 2013

e-Science

- Ok, mas por que e-Science é importante?



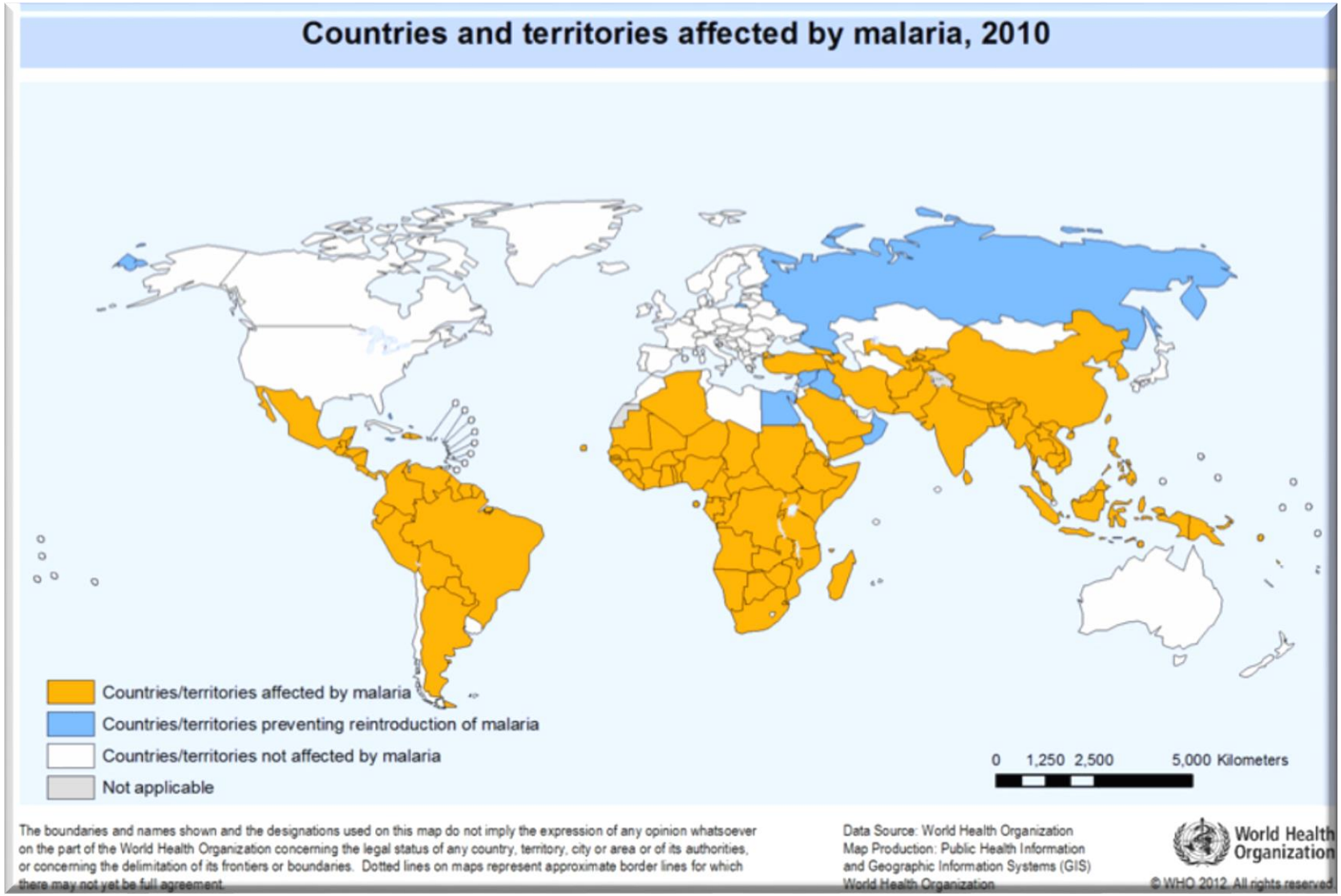
Experimento de Motivação

- DTN compreendem infecções parasitárias, virais e bacterianas que atacam especialmente as populações de baixa renda nas regiões em desenvolvimento da África, Ásia e América Latina
- 17 DTN em 149 países afetam 1 bilhão de pessoas (1/7 da população mundial)

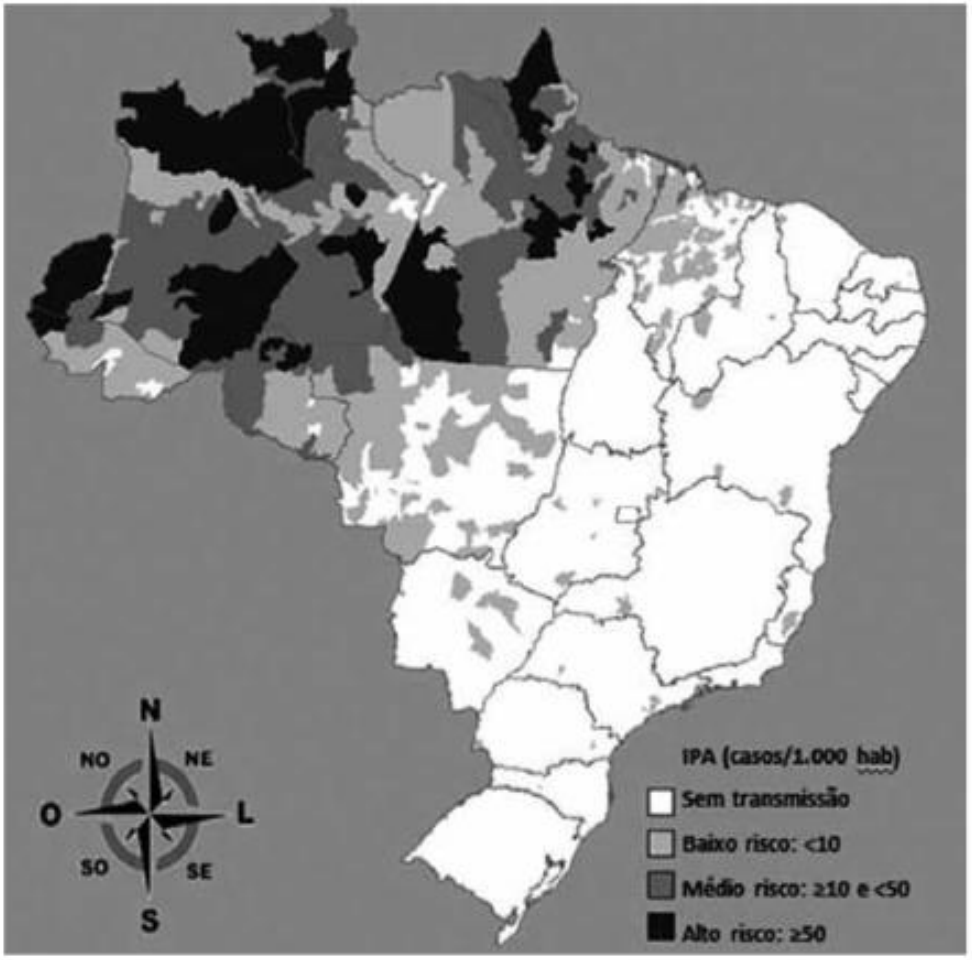
  - **Doença de Chagas**
  - **Leishmaniose**
  - **Hanseníase**
  - **Esquistossomose**
  - **Infecções por trematódeos transmitidas por alimentos**
  - **Helmintíases transmitida pelo solo**
  - **Cisticercose**
  - **Dengue**
  - **Dracunculíase**
  - **Equinococose**
  - **Oncocercose**
  - **Raiva**
  - **Tracoma**
  - **Úlcera de Buruli**
  - **Tripanossomíase humana africana**

- A OMS reporta progressos contra a **malária** como uma prioridade

Malária no Mundo



Malária no Brasil



Fonte: Sistema de Informação de Vigilância Epidemiológica-Malária/Secretaria de Vigilância em Saúde/Ministério da Saúde

Mapa de risco da malária por Município de Infecção. Brasil 2011.

Cisteíno Proteases em *Plasmodium*

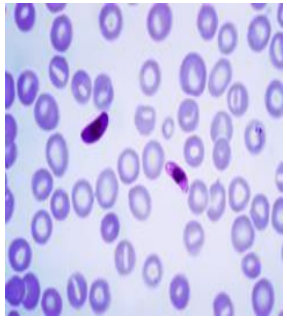
- *Plasmodium falciparum* gera os casos mais graves de malária
- As drogas antimaláricas apresentam uma crescente resistência
- Evidências indicam que **cisteíno proteases (CP)** têm um papel essencial no tratamento médico da malária
- **Projetos de descoberta de drogas**

  →alvos fármaco-terapêuticos promissores

  →inibidores antimaláricos

  →papel importante no ciclo de vida do parasita

  →papaína e ubiquitina



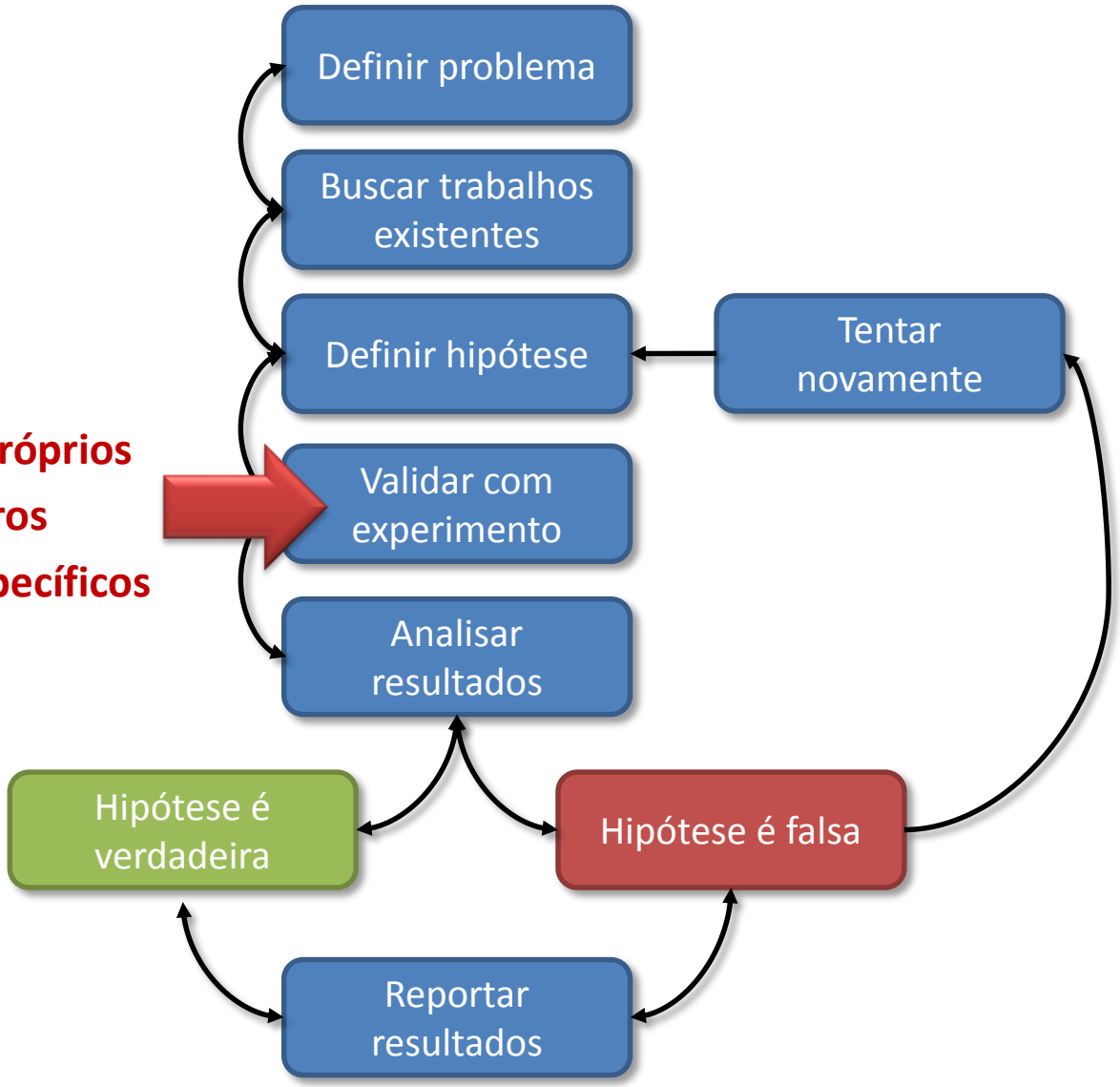
Plasmodium falciparum

Testes para Análise de Cisteíno Proteases

Cientistas desenvolvem códigos próprios

Cientistas usam códigos de terceiros

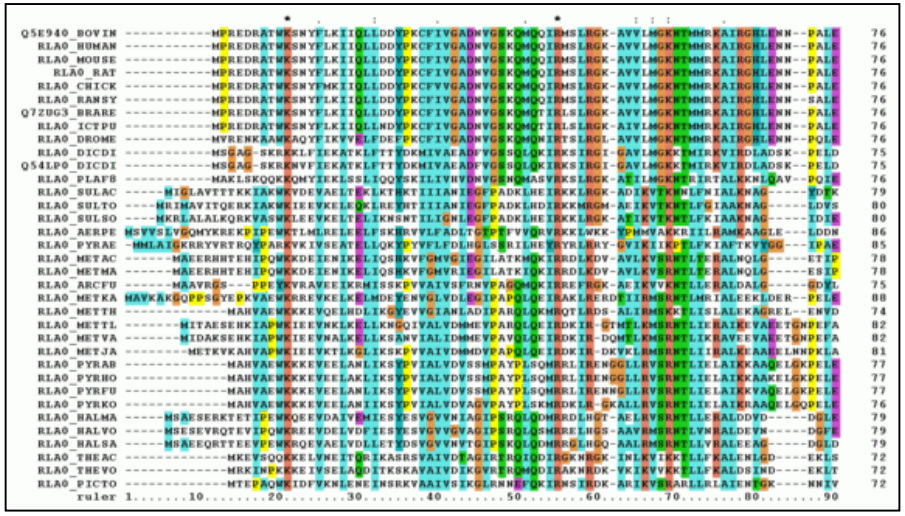
Desenvolvimento de pipelines específicos

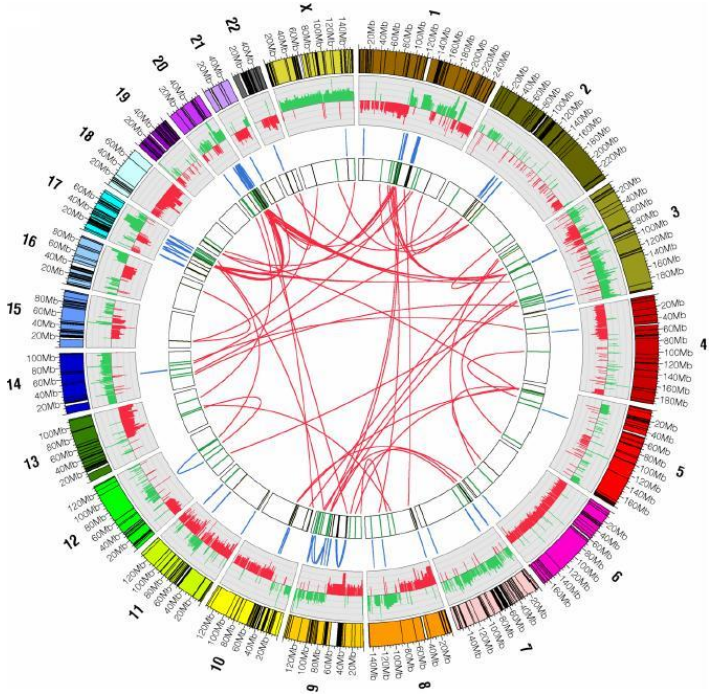


http://www.sciencebuddies.org

1º Passo - Genômica Comparativa

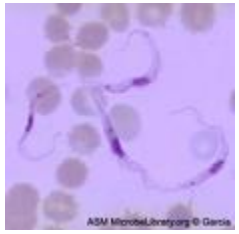
- Um alinhamento múltiplo de sequência (AMS) identifica regiões de similaridade que podem ser consequência de relações funcionais, estruturais ou evolutivas entre as sequências

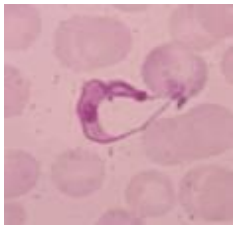




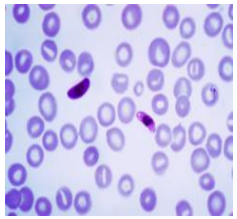
2º Passo - Filogênia

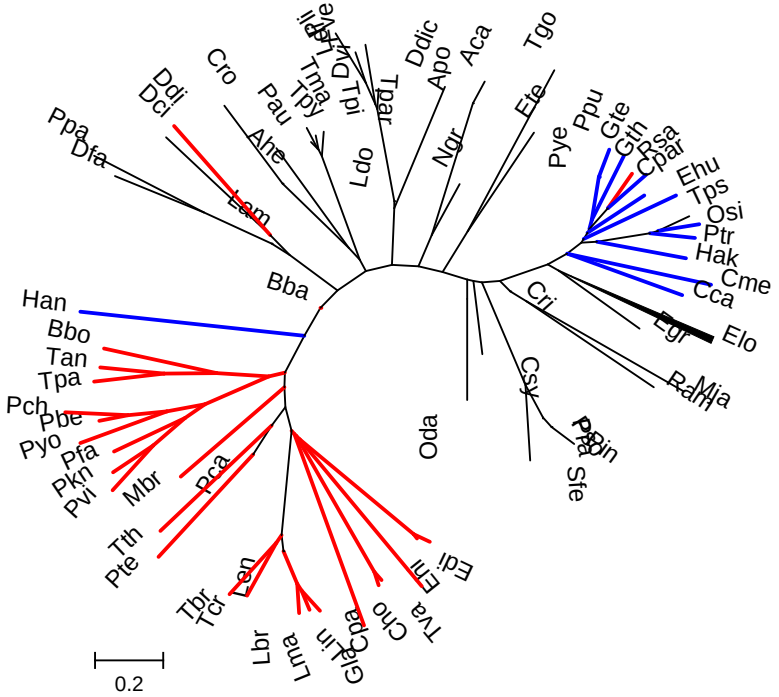
- Faz uso extensivo de AMS para a construção de árvores filogenéticas usados para inferir relações evolutivas entre genes homólogos


Trypanosoma cruzi


Leishmania major


Trypanosoma brucei

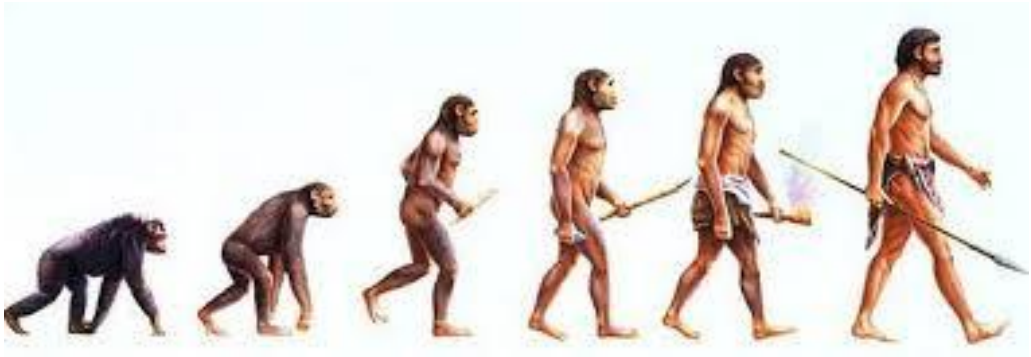

Plasmodium falciparum

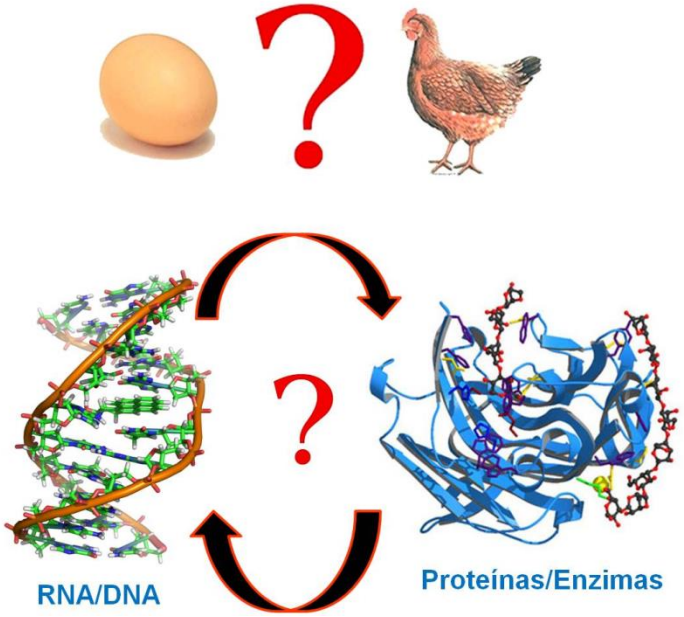


Ocaña K.A.C.S., and Dávila A.M.R. Phylogenomics-Based Reconstruction of Protozoan Species Tree. *Evolutionary Bioinformatics*, vol. 7, pp. 107-121, 2011.

3º Passo - Evolução Molecular

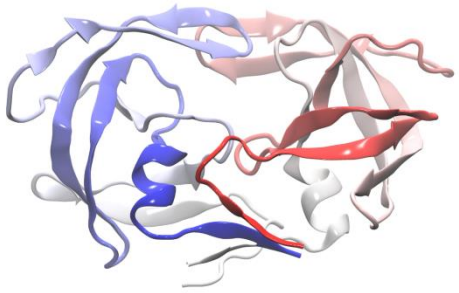
- Fornece base para inferências evolutivas e biológicas, e evoluem devido ao crescimento explosivo de dados genômicos e novas abordagens computacionais e estatísticas

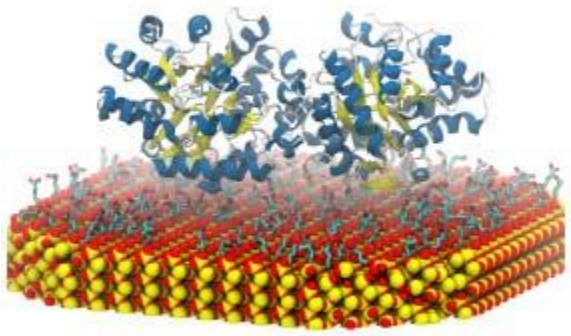


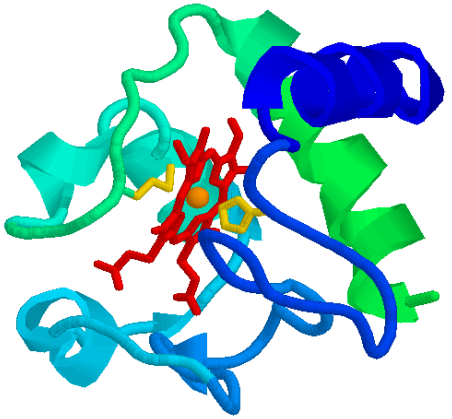


4º Passo - Modelagem Molecular

- Construção do modelo de resolução atômico (**modelo 3D**) da sequência da proteína (**alvo**) e uma estrutura 3D experimental de proteínas homólogas (**molde**)
- Uso de métodos de Genômica Estrutural *e.g.*, *Computer-Aided Drug Design* (CADD)



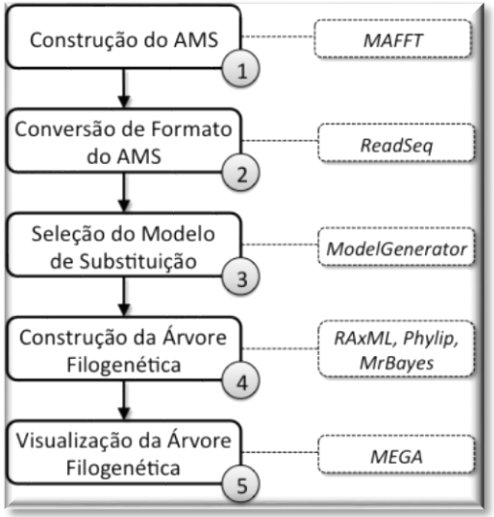




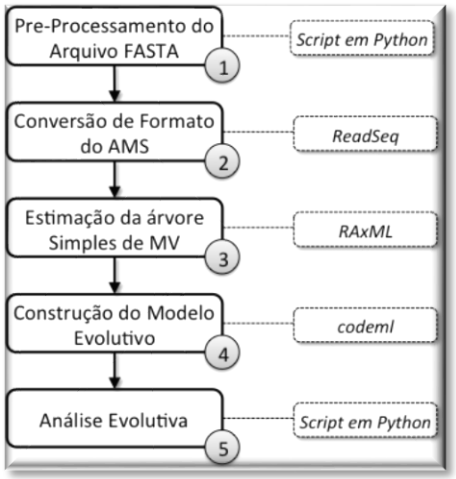
Pipelines Farmacofilogenômica

(A) Genômica Comparativa

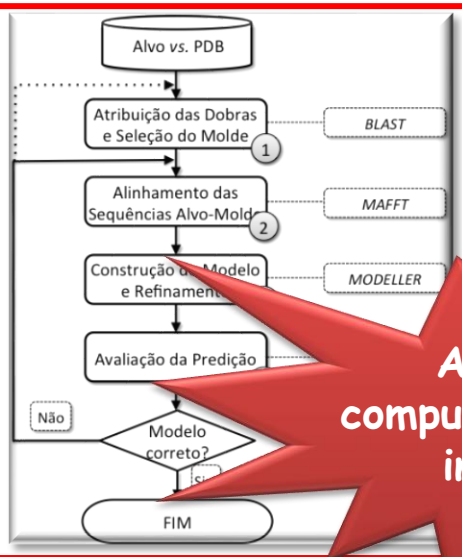
(B) Filogenia



(C) Evolução Molecular



(D) Modelagem Molecular



Aplicações computacionalmente intensivas

Pipelines Farmacofilogenômica

(A) Genômica Comparativa

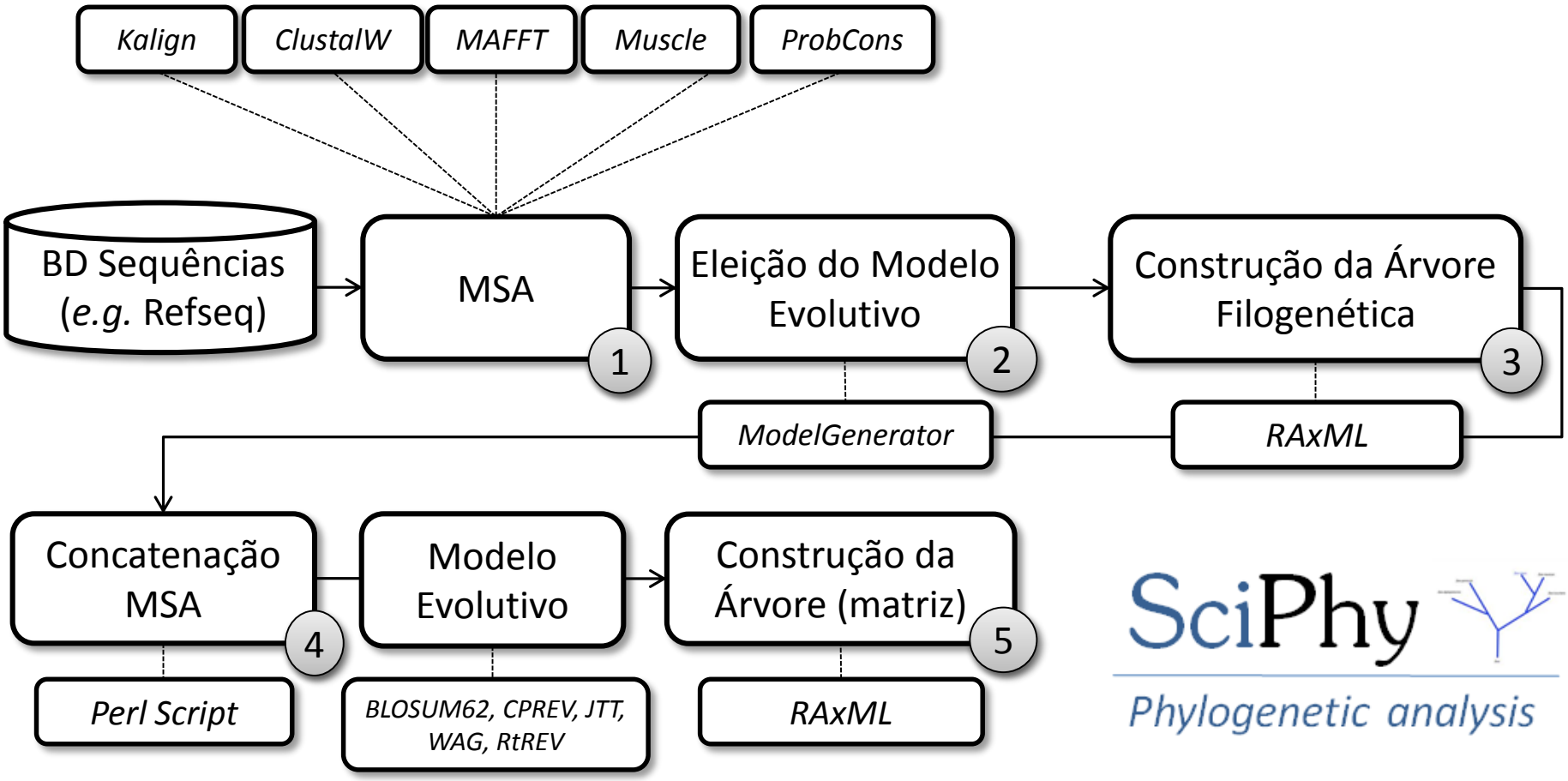
(B) Filogenia

(C) M… …gem …lar

Como gerenciar todas essas execuções?

Aplicações computacionalmente intensivas

Análise de *Cisteíno Proteases* Modelada como *Workflow* Científico



Fonte: Ocaña K.A.C.S., Oliveira D., Ogasawara E., Dávila A.M.R., Lima A.A.B., and Mattoso M. **SciPhy: A Cloud-Based Workflow for Phylogenetic Analysis of Drug Targets in Protozoan Genomes**. Springer, pp. 66-70, 2011.

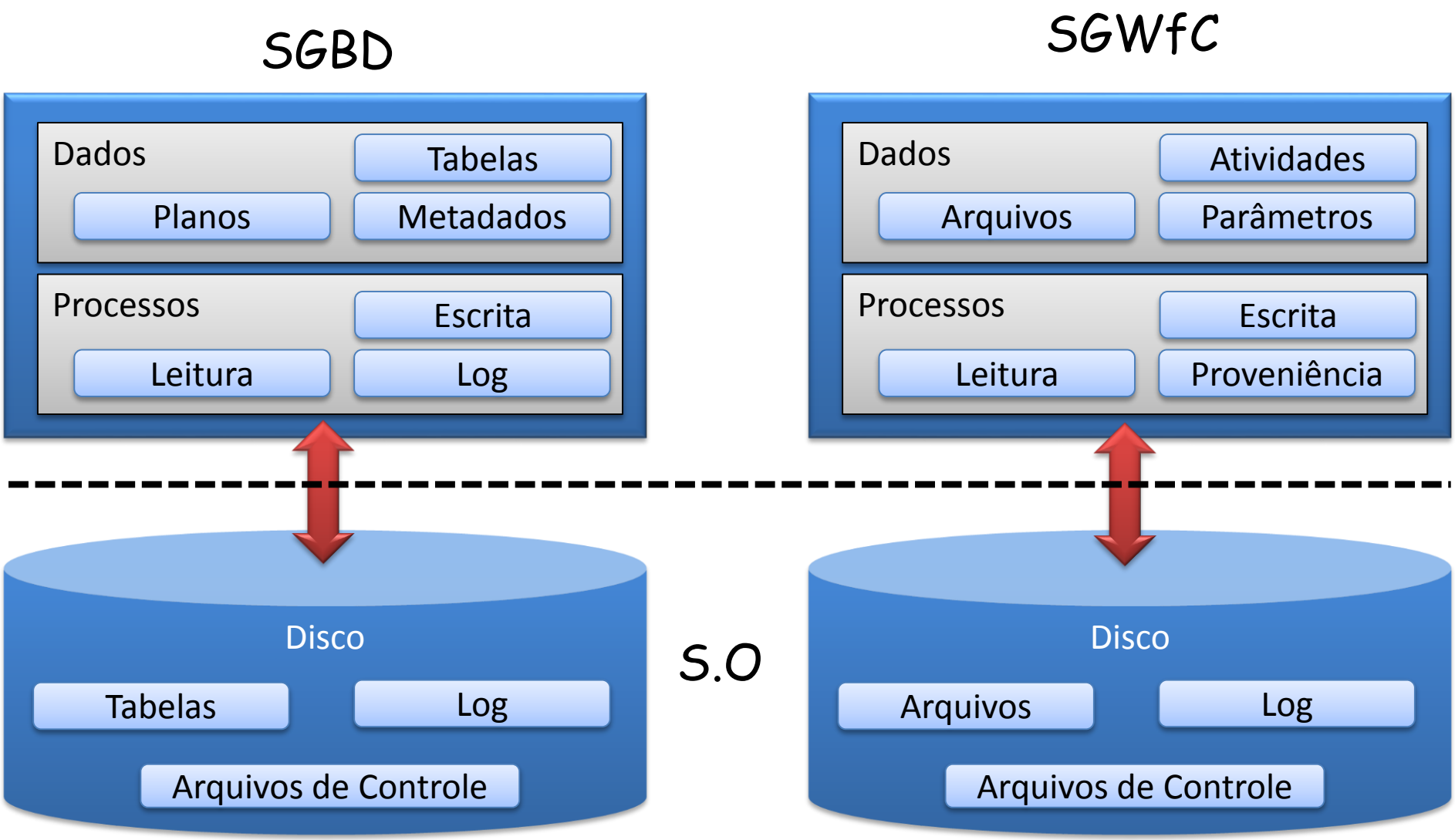
Workflows Científicos

- *Workflows* científicos são abstrações que representam a cadeia de atividades dentro de um experimento científico
- Eles são gerenciados por Sistemas de Gerência de *Workflows* Científicos (SGWfC)
- Diversas variações do *workflows* dentro do mesmo experimento
- Essas variações incluem a alteração de dados de entrada e parâmetros
- **Proveniência é uma questão fundamental**

Sistemas de Gerência de Workflows Científicos

- Definem e executam os workflows
- Proveem execução eficiente
- Controlam falhas garantindo a integridade
- Acessam/Armazenam/Consultam dados usando SGBD
- Rastreiam a **proveniência** dos dados

Analogia com Banco de Dados



Abordagens Existentes para Execução de *Workflows* Científicos

- Modelo de execução com paralelismo nativo
  Ex.: Swift, Pegasus, Triana, SciCumulus

- Motor sequencial
  Ex.: Kepler, VisTrails, Taverna

- Motor sequencial e camada de paralelismo
  Ex.: VisTrails+Hadoop, Kepler+Hadoop

Execução de uma Análise de *Cisteíno Protease*

- Execução demanda grande tempo de processamento
- Cientistas necessitam variar:
  - Parâmetros
  - Métodos
  - Imagens
  - Arquivos de Dados
  - Configurações
- Como executar essas análises?

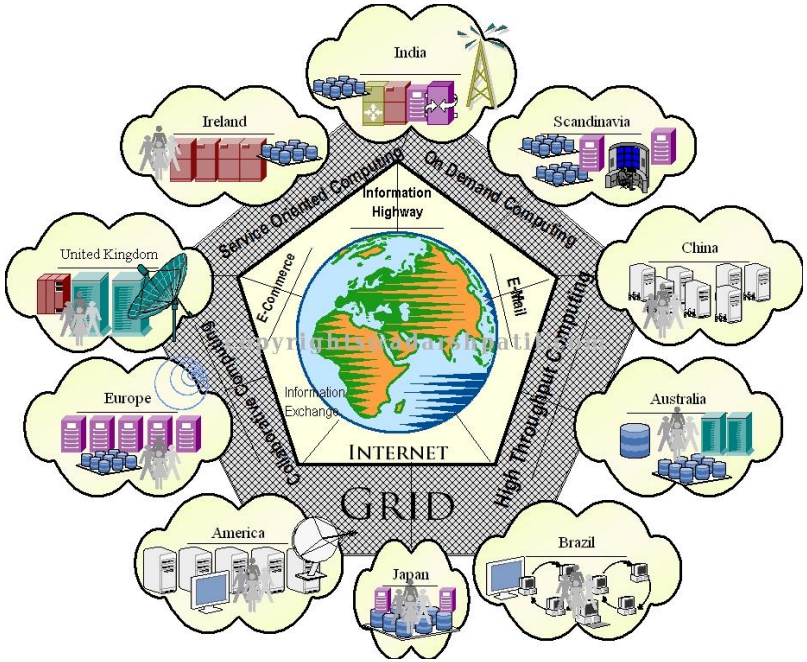
Execução de uma Análise de *Cisteíno Protease*

- Execução demanda grande tempo de processamento
- Cientistas necessitam variar:
  - Parâmetros
  - Métodos
  - Imagens
  - Arquivos de Dados
  - Configurações
- Como executar essas análises?

Paralelismo!

Em que ambiente executar?

Clusters



Grades

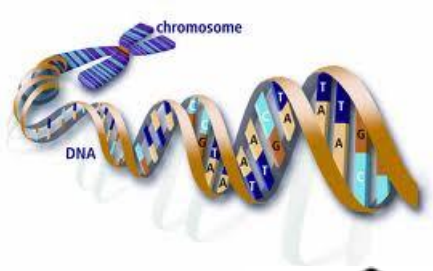
Desktop

Em que ambiente executar?

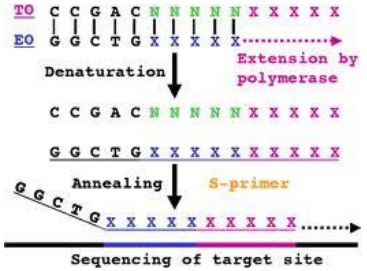
- Clusters podem ser financeiramente custosos
- Grades demandam um esforço grande de configuração
- Desktops podem demandar um tempo exageradamente grande para a execução
- **Tempo de execução *versus* Tempo de Análise dos Dados**

Alternativa: Análise de *Cisteíno Protease* em Nuvens

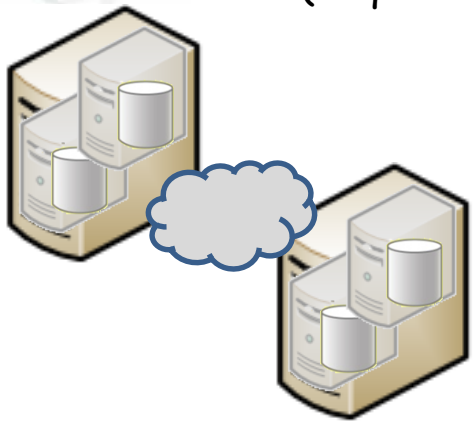
Dados de entrada para Realizar análises filogenéticas: Sequências de DNA e RNA de diversos organismos...



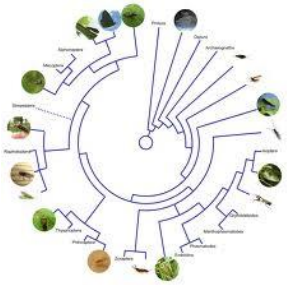
1. Alinhamento de sequências (MAFFT, ProbCons, ClustalW, Muscle e Kalign) e conversão



Produz uma grande quantidade de dados ... (sequências alinhadas)...

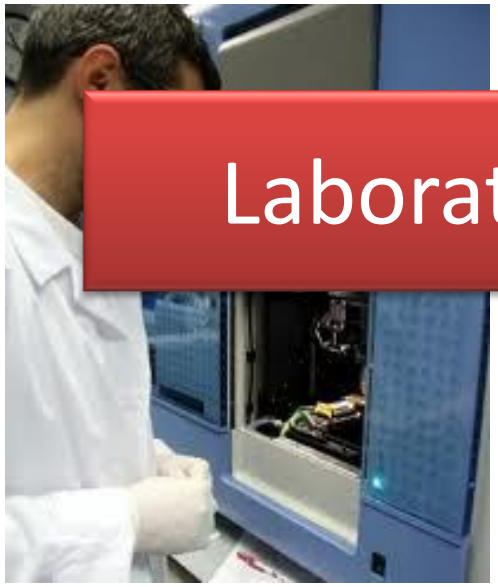


2. ... Utilizada para a geração do modelo evolutivo (ModelGenerator)



3. Geração das árvores filogenéticas (RAxML)

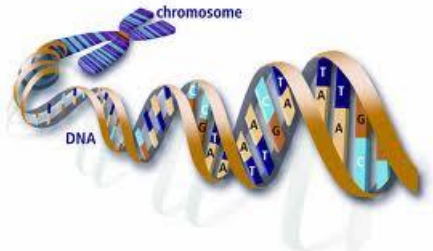
Alternativa: Análise de *Cisteíno Protease* em Nuvens



Laboratório

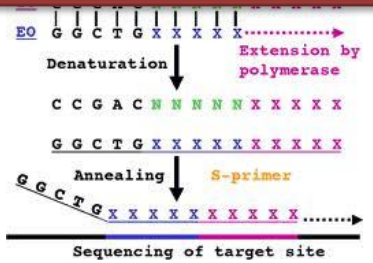
Dados de entrada para
ses

e DNA e
sos
organismos...



1. A
(MA
Mus
as
V,
ão

Desktop



Produz uma grande quantidade de dados ... (sequências alinhadas)...

Nuvem

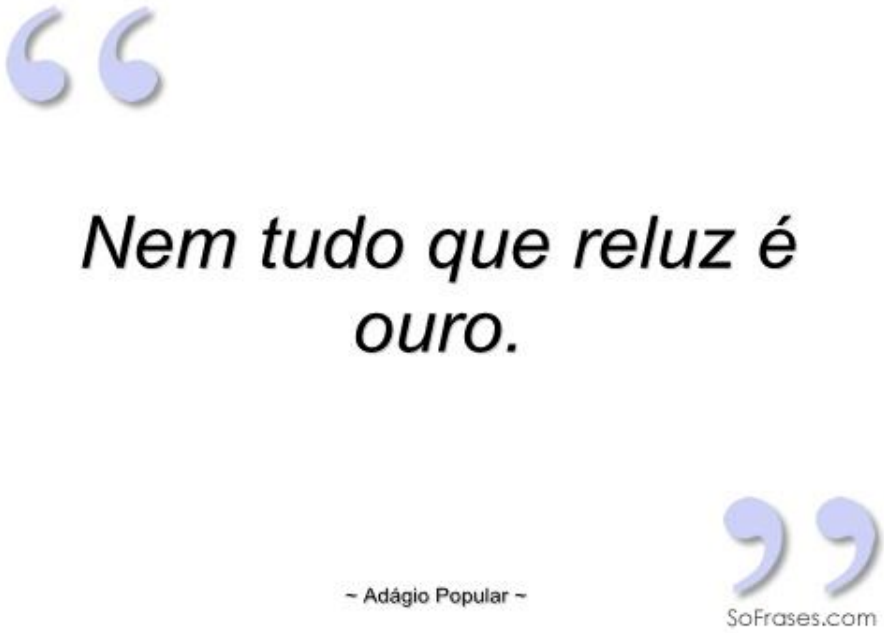
2. ... Utilizada para a geração do modelo evolutivo (ModelGenerator)

Tiled Wall

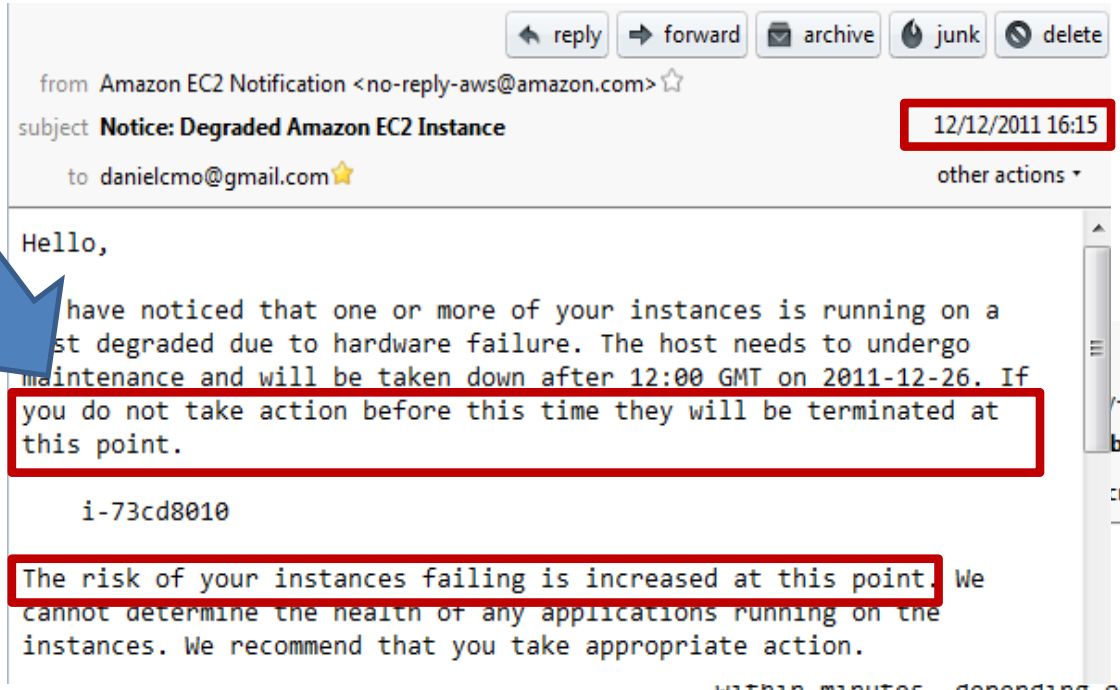
o das
s
icas
L)

Execução de uma Análise de Cisteíno Protease em Nuvens

- Utilização da nuvem faz com que cientistas não tenham que comprar equipamentos
- Execução paralela reduz o tempo de processamento
- Mas.....



Problemas na Execução de *Workflows* em Nuvens


reply forward archive junk delete

from Amazon EC2 Notification <no-reply-aws@amazon.com>

subject **Notice: Degraded Amazon EC2 Instance** 12/12/2011 16:15

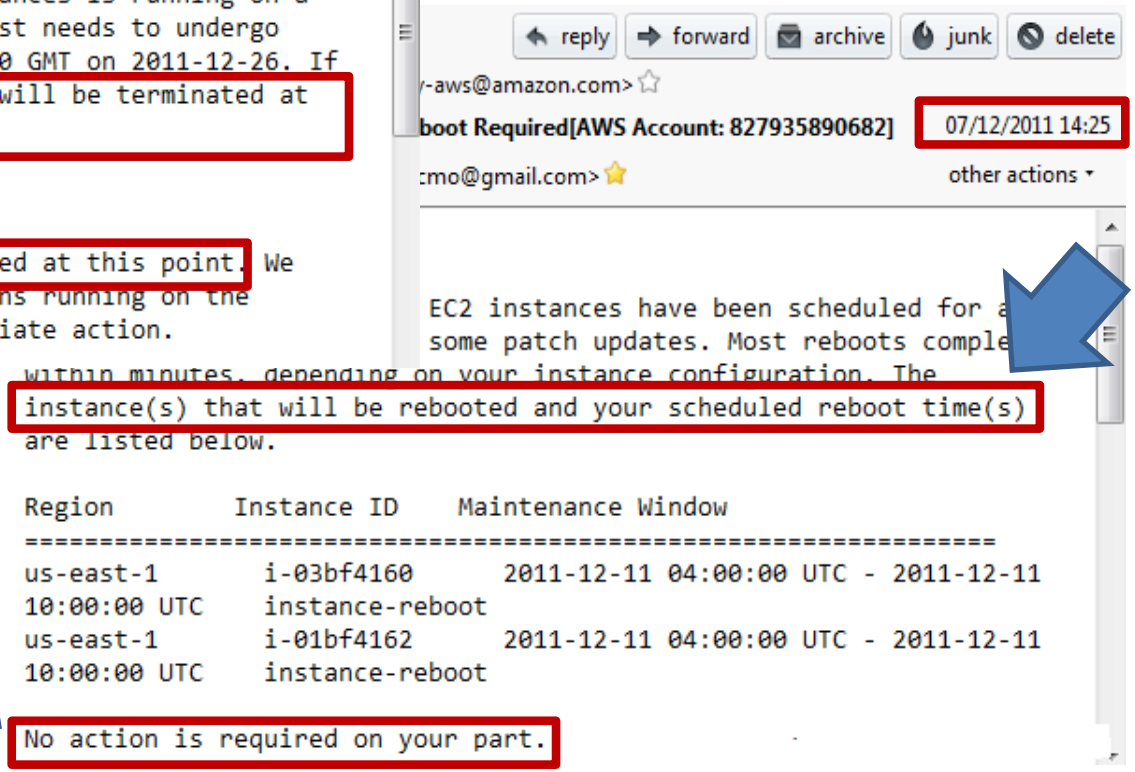
to danielcmo@gmail.com other actions

Hello,

I have noticed that one or more of your instances is running on a host degraded due to hardware failure. The host needs to undergo maintenance and will be taken down after 12:00 GMT on 2011-12-26. If you do not take action before this time they will be terminated at this point.

i-73cd8010

The risk of your instances failing is increased at this point. We cannot determine the health of any applications running on the instances. We recommend that you take appropriate action.


reply forward archive junk delete

-aws@amazon.com>

boot Required[AWS Account: 827935890682] 07/12/2011 14:25

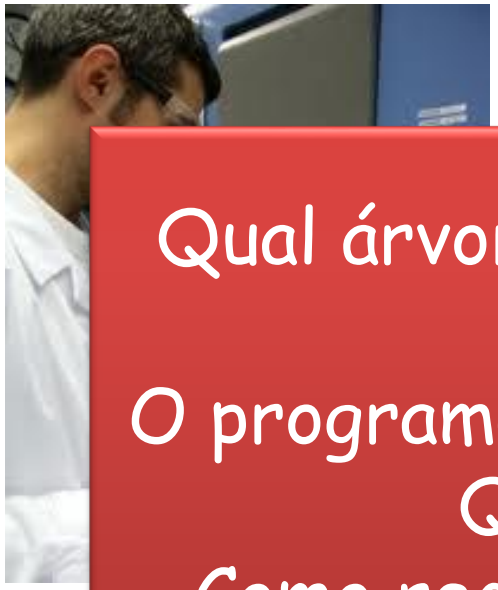
cmo@gmail.com> other actions

EC2 instances have been scheduled for a some patch updates. Most reboots comple within minutes, depending on your instance configuration. The instance(s) that will be rebooted and your scheduled reboot time(s) are listed below.

Region	Instance ID	Maintenance Window
us-east-1	i-03bf4160	2011-12-11 04:00:00 UTC - 2011-12-11 10:00:00 UTC
	instance-reboot	
us-east-1	i-01bf4162	2011-12-11 04:00:00 UTC - 2011-12-11 10:00:00 UTC
	instance-reboot	

No action is required on your part.

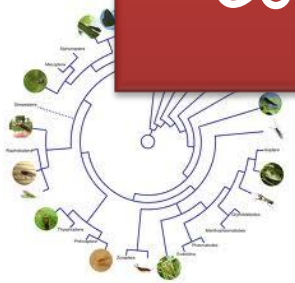
Nova Visão: Análise de Cisteíno Protease em Nuvens



Dados de entrada para Realizar análises

1. Alinhamento de sequências (MAFFT, ProbCons, ClustalW, Muscle e Kalign) e conversão

Qual árvore corresponde a uma determinada sequência de entrada?
O programa MAFFT é melhor que o ProbCons?
Qual e-value foi escolhido?
Como recuperar uma execução que falhou?
Como monitorar uma execução a distância?



árvores filogenéticas (RAxML)

ra a geração do modelo evolutivo (ModelGenerator)

Execução de uma Análise de de Cisteíno Protease

- Execução paralela reduz o tempo de processamento
- Utilização da nuvem faz com que cientistas não

Como relacionar todos esses recursos?

  - Arquivos de Dados
  - Configurações

Importância da Proveniência na Ciência

- Interpretar e reproduzir dados
- Verificar se um determinado experimento foi executado seguindo os procedimentos
- Identificar as entradas e saídas dos experimentos
- Garantir qualidade dos dados
- Rastrear quem executou um experimento e quem é responsável pelos resultados

"Provenance is as (or more) important as the results"
(Juliana Freire e Susan Davidson, SIGMOD 2008)

Desafios na Execução de *Workflows* em Nuvens

- Ambiente "caixa-preta" e em constante mudança
  - Variação de desempenho das máquinas virtuais
  - Mudanças constantes no ambiente
  - Falhas constantes nas máquinas virtuais
- Diversos tipos de máquinas virtuais passíveis de utilização
  - Qual tipo utilizar?
- Transferência de dados
  - Como evitar sobrecarga na transferência?
- Como monitorar os *workflows*?

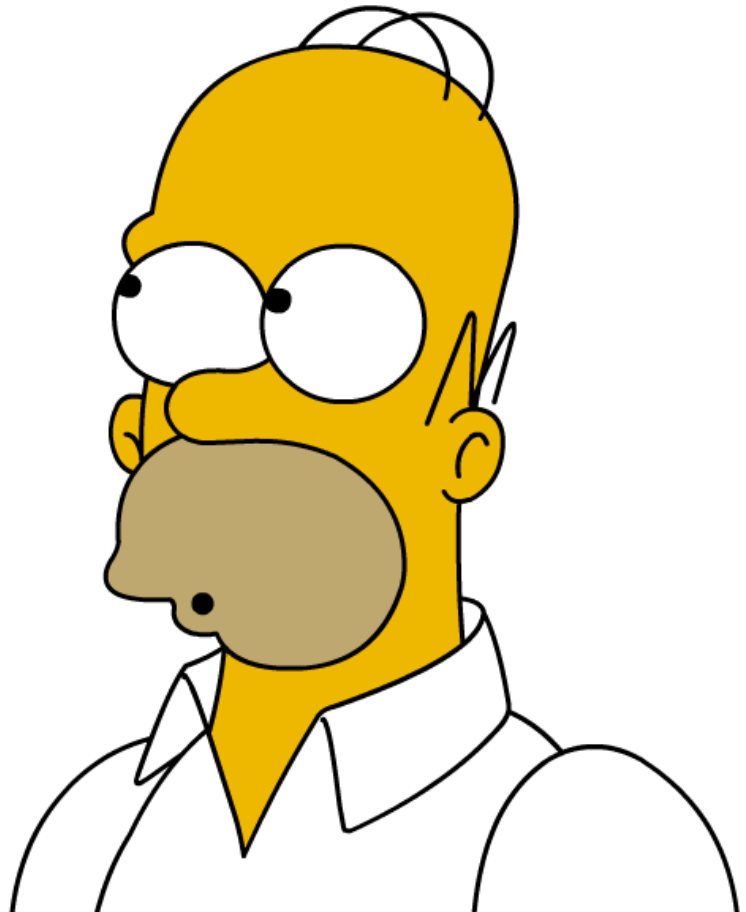
Questão

- "...considerando os desafios na execução de *workflows* em nuvem é possível prover **capacidade de alto desempenho** e a **gerência distribuída do *workflow* científico** por meio da adoção de soluções que considerem as **características únicas do ambiente, distribuam** as atividades e **monitorem** a execução das mesmas em várias máquinas virtuais?"

Hipótese

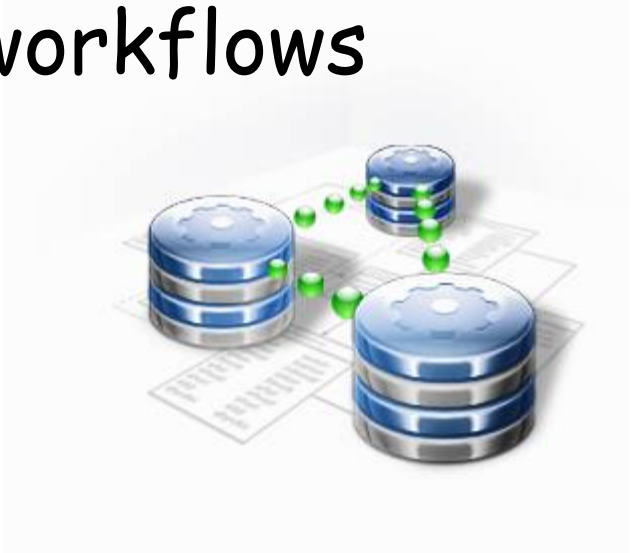
- "***Se*** *for adotada uma infraestrutura adaptativa para o escalonamento, despacho, monitoramento e captura de proveniência distribuída das atividades paralelas de um workflow científico* ***então*** *podemos prover a capacidade de alto desempenho, monitoramento e controle que são necessárias aos cientistas."*

Em que podemos nos inspirar na área de Banco de Dados?



Em que podemos nos inspirar na área de Banco de Dados?

- Representação da consulta por meio de álgebra relacional ≅ representação do workflows por meio de álgebra de workflows
- Otimização do plano de execução de consultas ≅ otimização de workflows científicos



Em que podemos nos inspirar na área de Banco de Dados?

- Escalonamento das atividades no ambiente de nuvem de acordo com uma álgebra de workflows
- Distribuição dos dados de proveniência
- Monitoramento

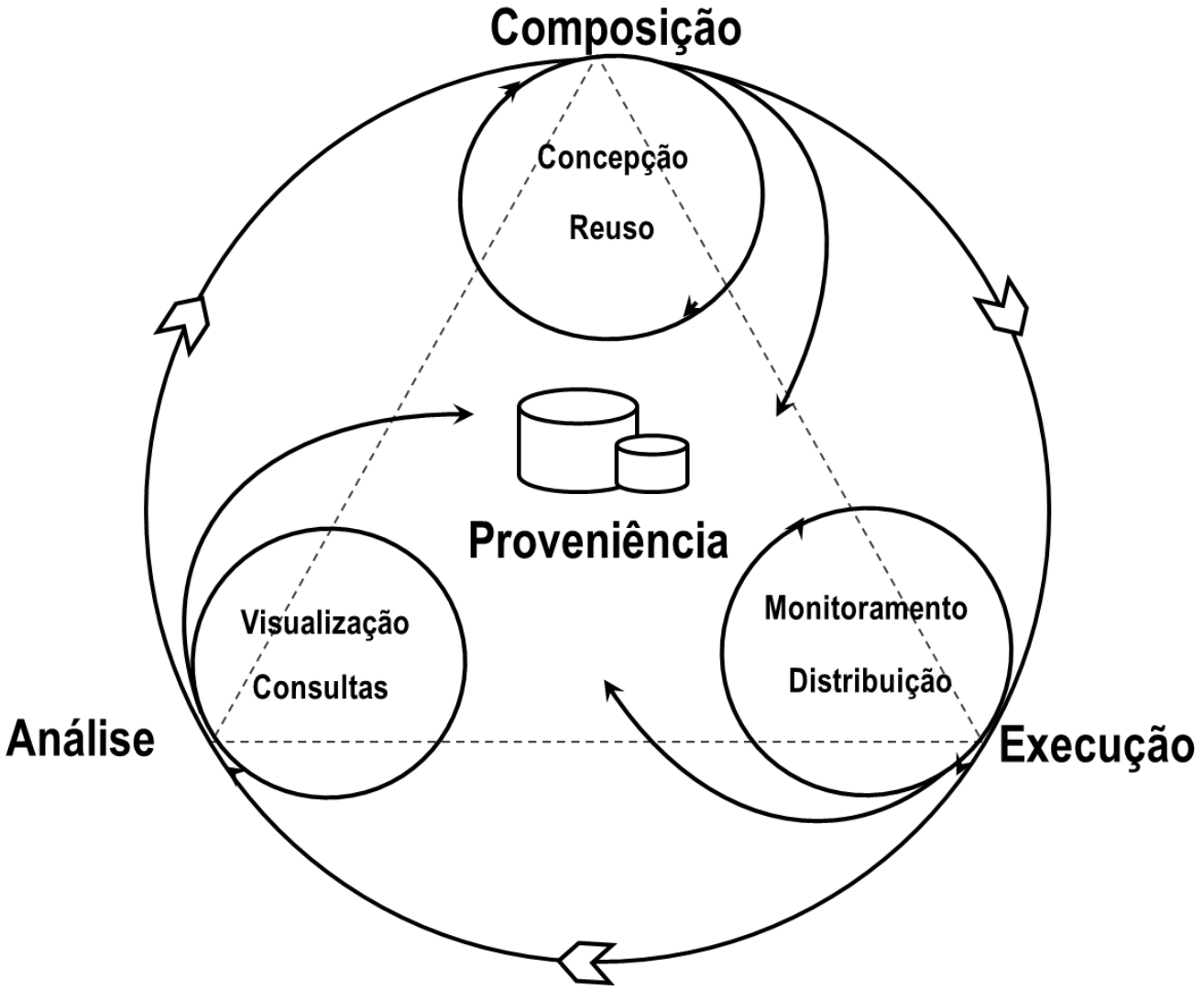
Roteiro do Tutorial

- Motivação
- **Workflows Científicos**
- Nuvens de Computadores
- Proveniência de Dados
- Máquinas de Workflow para Nuvem
- Aplicação de Proveniência em e-Science
- Demo

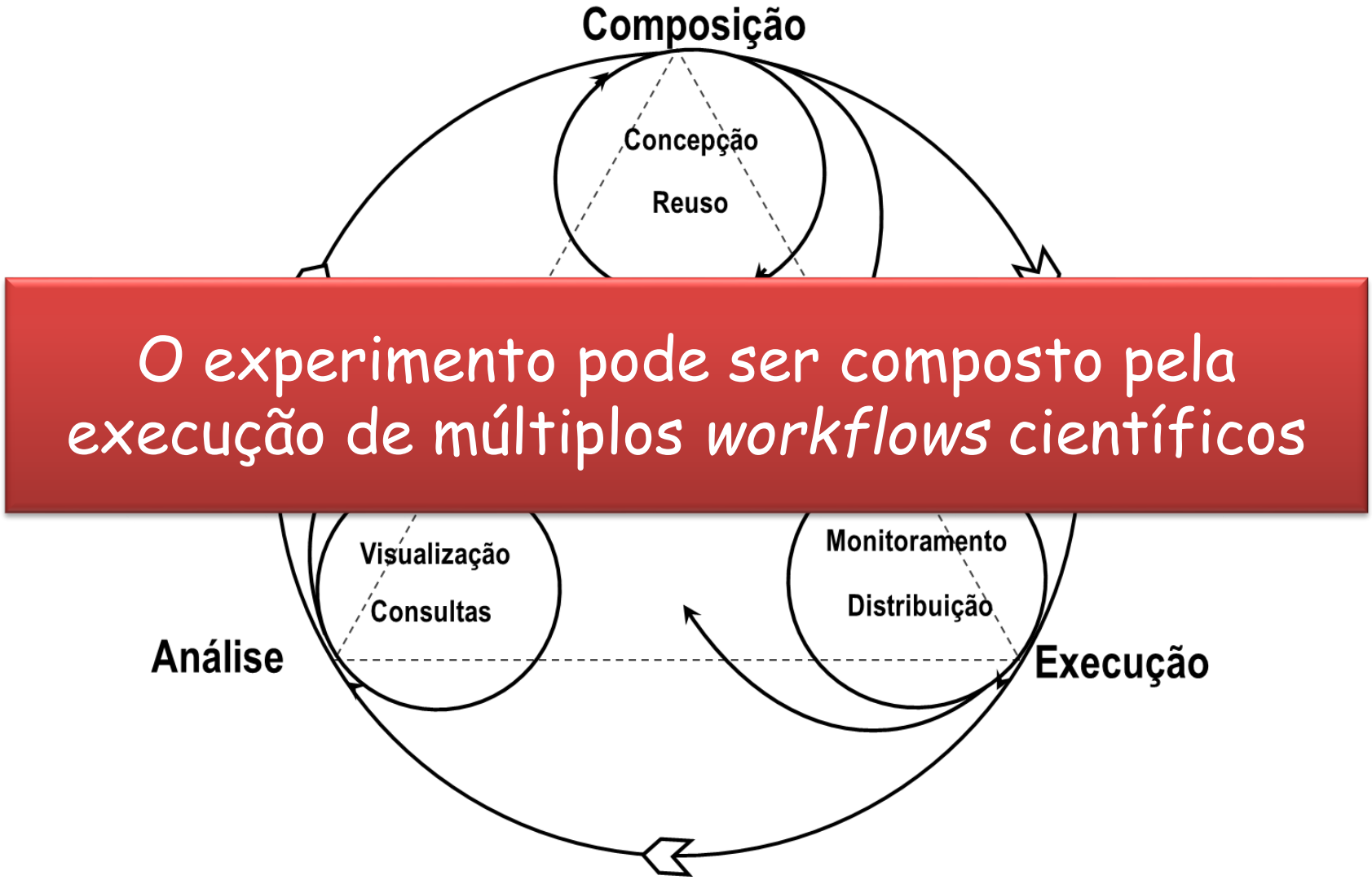
Experimento Científico

- Uma das formas utilizadas pelos cientistas para apoiar a formulação de novas teorias
- Possui um ciclo de vida com três etapas:
  - Composição
  - Execução
  - Análise

Nossa visão do ciclo de vida do experimento científico



Nossa visão do ciclo de vida do experimento científico



O experimento pode ser composto pela execução de múltiplos *workflows* científicos

Workflow

A automação de um processo de negócio, completo ou apenas parte dele, através do qual, documentos, informações ou tarefas são transmitidas de um participante a outro por ações, de acordo com regras procedimentais. (WFMC, 1995)

- Provê a abstração necessária para descrever uma série de processos estruturados e suas atividades, oferecem um contexto robusto de resolução de problemas e promovem o uso efetivo e otimizado dos recursos computacionais

Workflow

- Uma coleção de atividades organizadas para acompanhar algum experimento (processo de negócio).
- As atividades ou tarefas são os componentes de software independentes que implementam alguma funcionalidade e são executadas por um ou mais sistemas de softwares.
- Exemplos de atividades incluem executar um programa, transformar um arquivo ou atualizar um banco de dados.

Workflow

- Um workflow define a ordem de execução dessas atividades ou as condições em que essas atividades serão executadas e a sua eventual sincronização.
- Os dados de entrada e saída das atividades (variáveis) são definidos como o fluxo de dados do workflow.

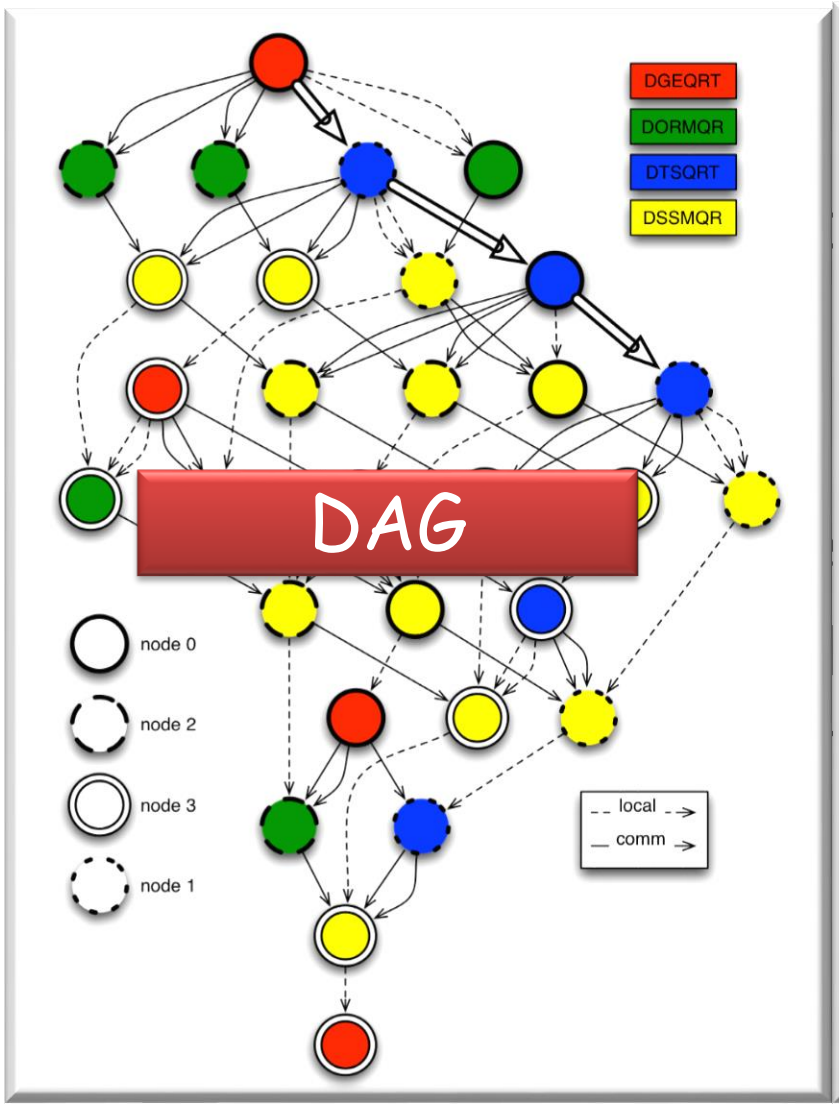
SGWf e Máquina de Workflow

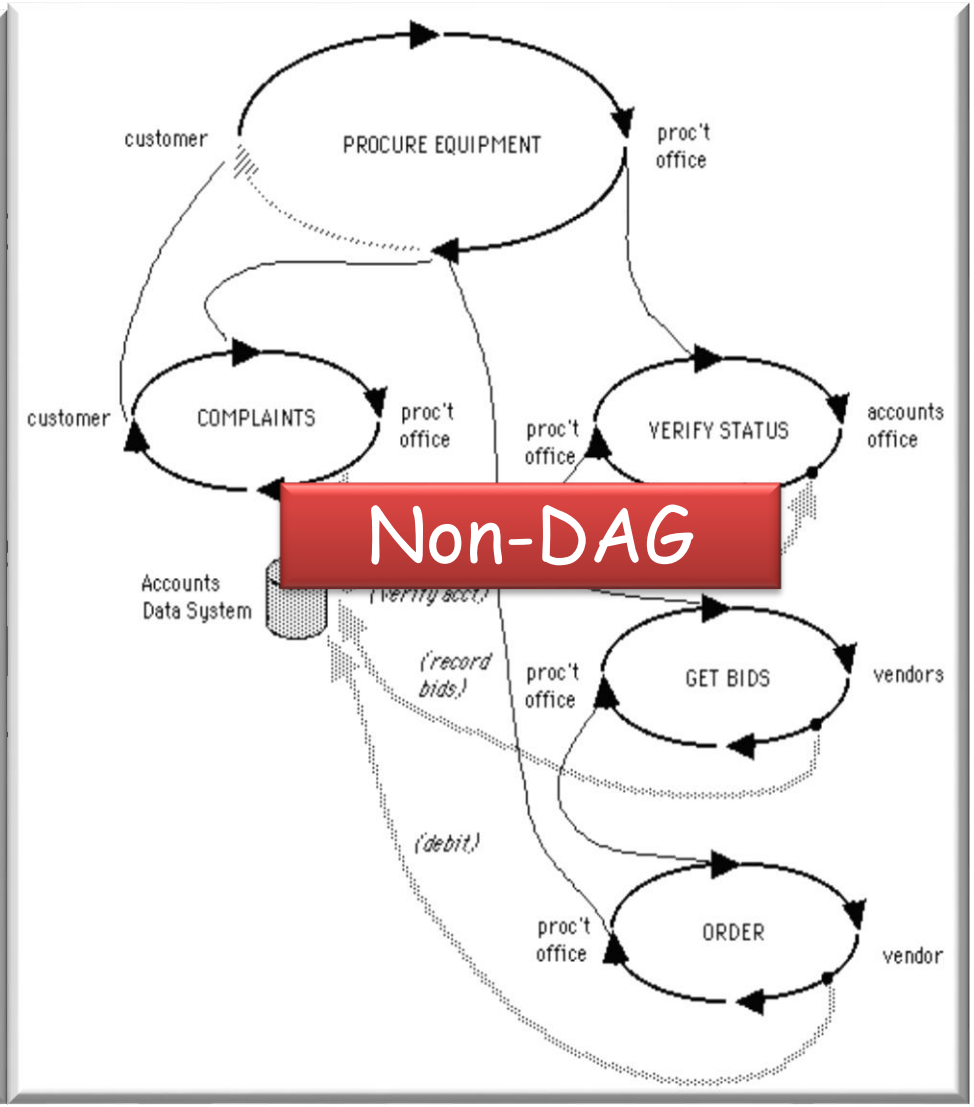
- Sistema de gerência de workflows – SGWf (WfMS - Workflow Management Systems)
  - Software que fornece toda a infraestrutura para definir, executar e monitorar workflows.
- Máquina de Workflow
- Na utilização deste tipo de software é importante para separar a definição do workflow, da máquina (*i.e. engine*) encarregada de executar o mesmo

Especificação do Workflow

- DAG
  - Contém estruturas do tipo sequenciais, paralelas ou livres.
- Non-DAG
  - Incluem estruturas de iteração
  - Permitem que os workflow executem tarefas ou sub-workflows de forma repetida.

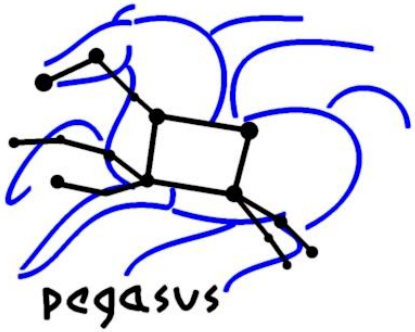
Especificação do Workflow





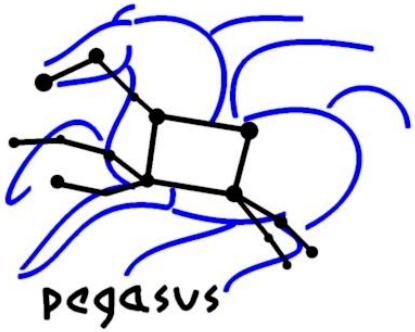
Diversidade de Sistemas de Gerência de Workflows

- Askalon
- Chiron
- Kepler
- Pegasus
- SciCumulus
- Swift
- Taverna
- VisTrails



Diversidade de Sistemas de Gerência de Workflows

- Askalon
- **Chiron**
- Kepler
- Pegasus
- **SciCumulus**
- Swift
- Taverna
- VisTrails



SGWf...um caldeirão de tecnologias

- Muitos Modelos…
  - Computação – control flow, data flow, pipelines baseados em scripts
  - Interação– interativos, batch
  - Representação de Wf – DAG, non-DAG
  - Adaptabilidade – dinâmicos, estáticos
- Muitos Ambientes…
  - Componentes – grid, cluster, **nuvem**, etc
  - Natureza- open-source, comerciais
  - Escala– long running,data streaming, etc…
- Muitas tecnologias….
  - centralizado, distribuído, local, Web…

Roteiro do Tutorial

- Motivação
- Workflows Científicos
- **Nuvens de Computadores**
- Proveniência de Dados
- Máquinas de Workflow para Nuvem
- Aplicação de Proveniência em e-Science
- Demo

Computação em Nuvem

- Uma das opções para prover um ambiente de alto desempenho para cientistas é executar seus experimentos em nuvens de computadores

- Se baseia na ideia de prover recursos (*software* e *hardware*) utilizando o ambiente *Web* para uma gama (virtualmente infinita) de usuários
  - Elasticidade de Recursos
  - Virtualização de Recursos
  - Personalização
  - Disponibilidade

Quem já usa Computação em Nuvem?

52% das grandes empresas já utilizam cloud computing, diz Nicholas Carr

75% das grandes empresas no Brasil já usam cloud computing, aponta estudo

Uso de computação na nuvem cresce 40% em 12 meses no Brasil

FAPEMIG lança edital para *cloud computing*
30 de Dezembro de 2011

Quem já usa Computação em Nuvem?


Home > Tecnologia > Cloud veio para ficar, diz IDC

Cloud veio para ficar, diz IDC


Virtualização e cloud computing são saídas para reduzir gastos

19/10/2010 9:15, Marcelo Bernstein, com agências - de Nova York


TI tem investido em cloud e virtualização, não em PCs

Embora o mercado de computadores tenha crescido 2,6%, o número é menor que a previsão inicial de 2,9%, aponta a IDC.

Shane O'Neill, CIO/EUA

Publicada em 15 de julho de 2011 às 12h01


Brasil ainda está na primeira fase de adoção da nuvem

Convergência Digital - Hotsite Cloud Computing

:: Por Fábio Barros* :: 28/07/2011


TI | 05/10/2011 14:06 » Compartilhar: +1 + redes

Computação em nuvem poderá movimentar US$ 1,45 bi em 2015

Segundo a empresa de pesquisa de mercado Gartner, a receita das empresas de serviços de armazenamento aumentou 56% em 2011


Cloud computing reduzirá consumo de energia de data centers em 31% até 2020

Convergência Digital - Hotsite Cloud Computing

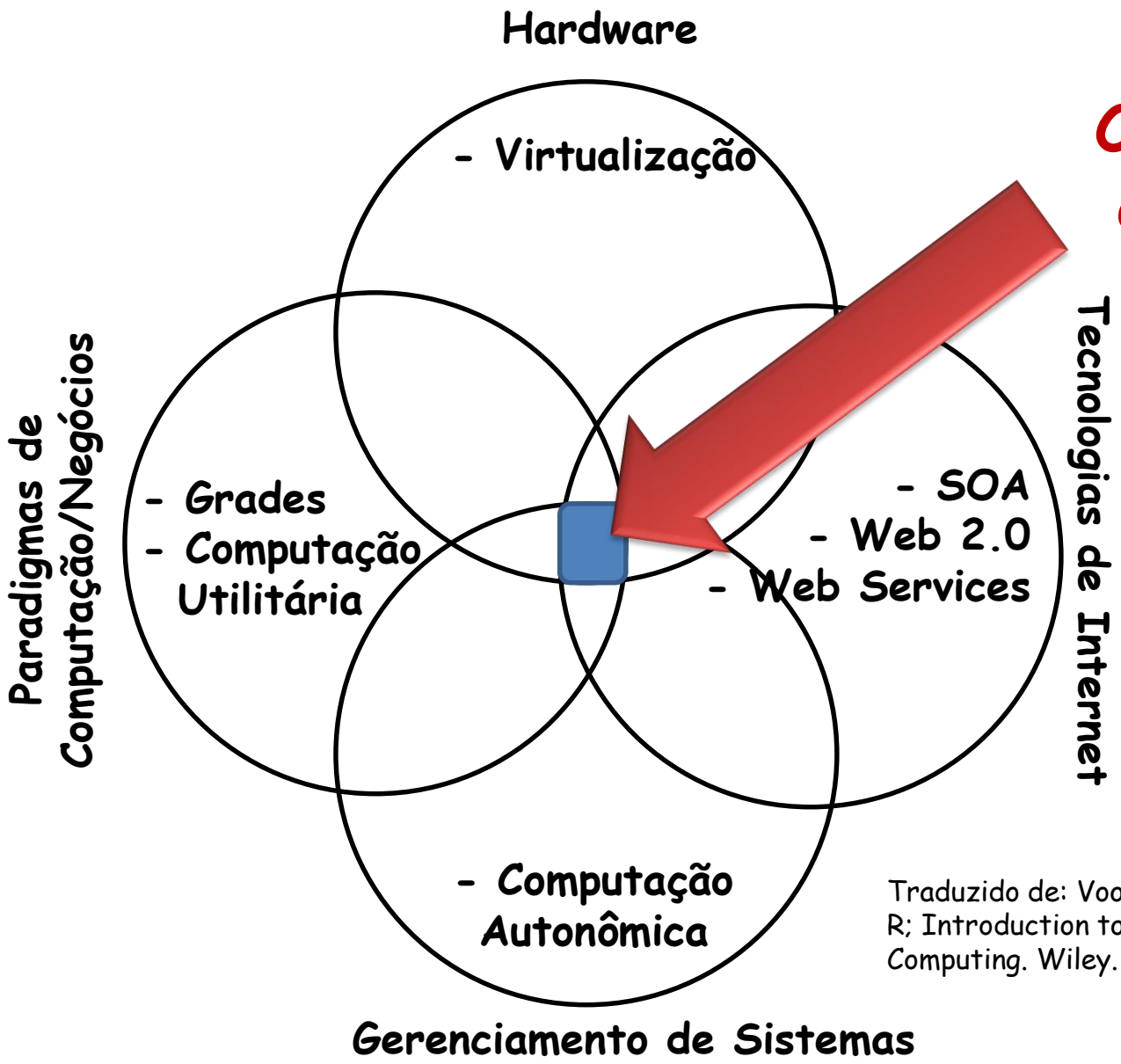
:: Da redação :: 21/09/2011

Computação em Nuvem: Definição

- Diversas definições: **não há consenso!**
- Mais de 20 definições em **Vaquero (2009)**
- Segundo **Ian Foster**:
  - "...um paradigma de computação em larga escala que possui foco em proporcionar **economia de escala**, em que um conjunto **abstrato, virtualizado, dinamicamente escalável** de poder de processamento, armazenamento, plataformas e serviços são disponibilizados **sob demanda** para clientes externos através da **Internet**."

Vaquero, L. M., Rodero-Merino, L., Caceres, J., and Lindner, M. (2009). A break in the clouds: towards a cloud definition. SIGCOMM Comput. Commun. Rev., 39(1):50–55

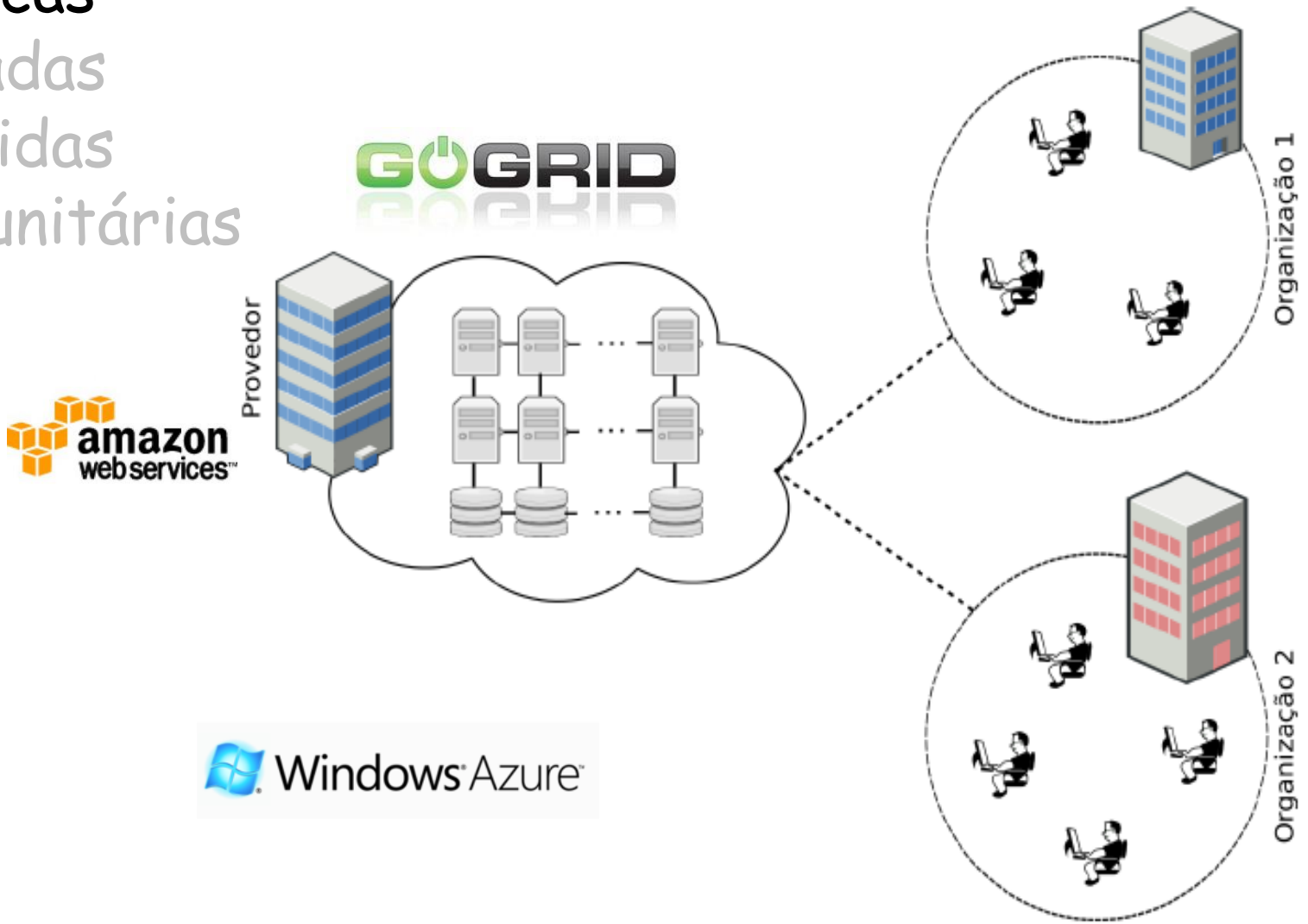
Computação em Nuvem



Traduzido de: Voorsluys, W; Broberg, J; Buyya, R; Introduction to Cloud Computing. IN: Cloud Computing. Wiley. 2011.

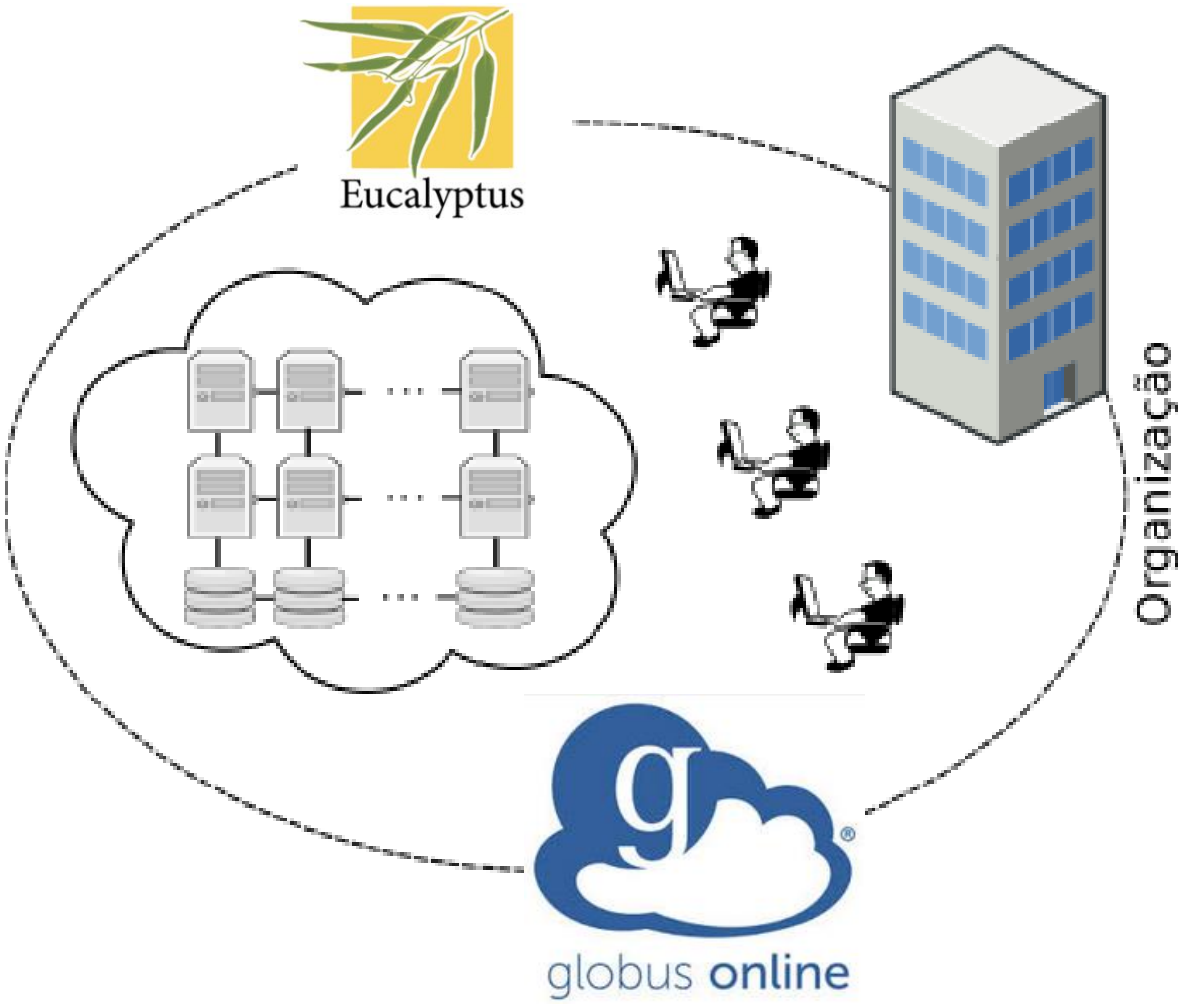
Tipos de Nuvem

- Públicas
- Privadas
- Híbridas
- Comunitárias



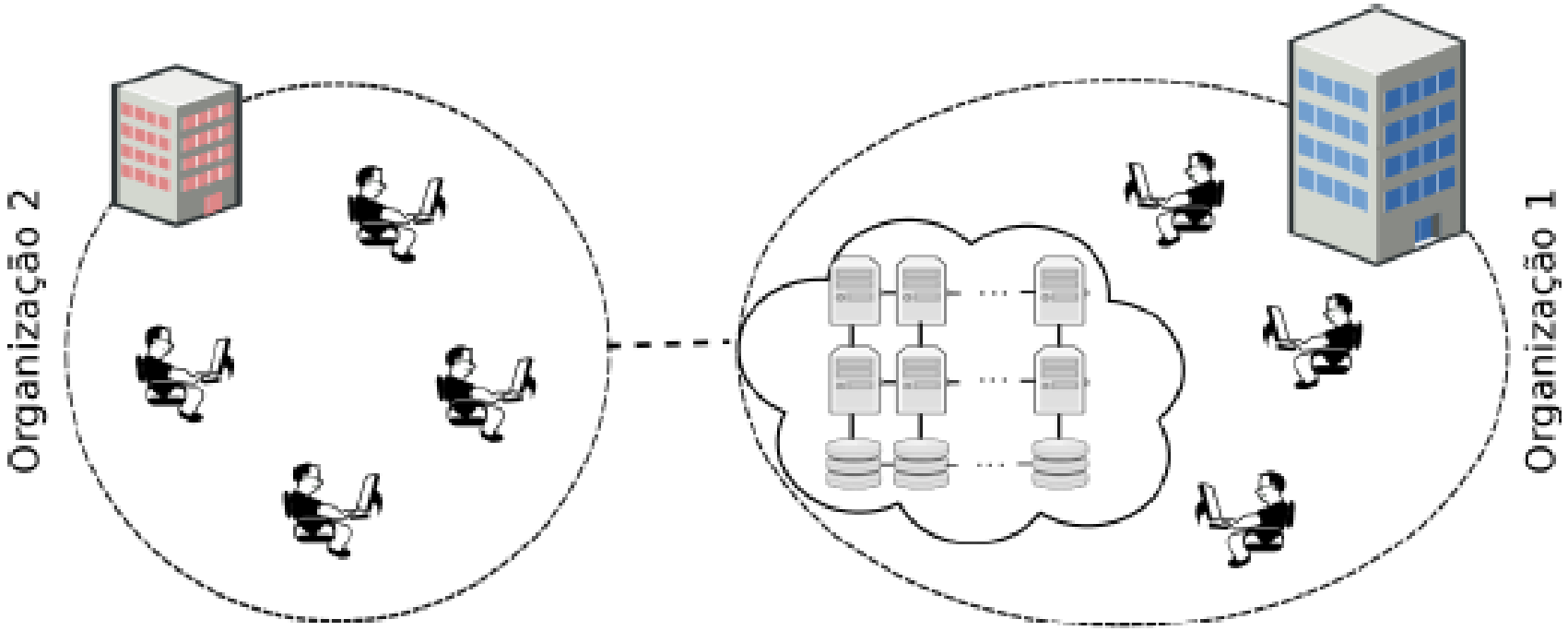
Tipos de Nuvem

- Públicas
- Privadas
- Híbridas
- Comunitárias



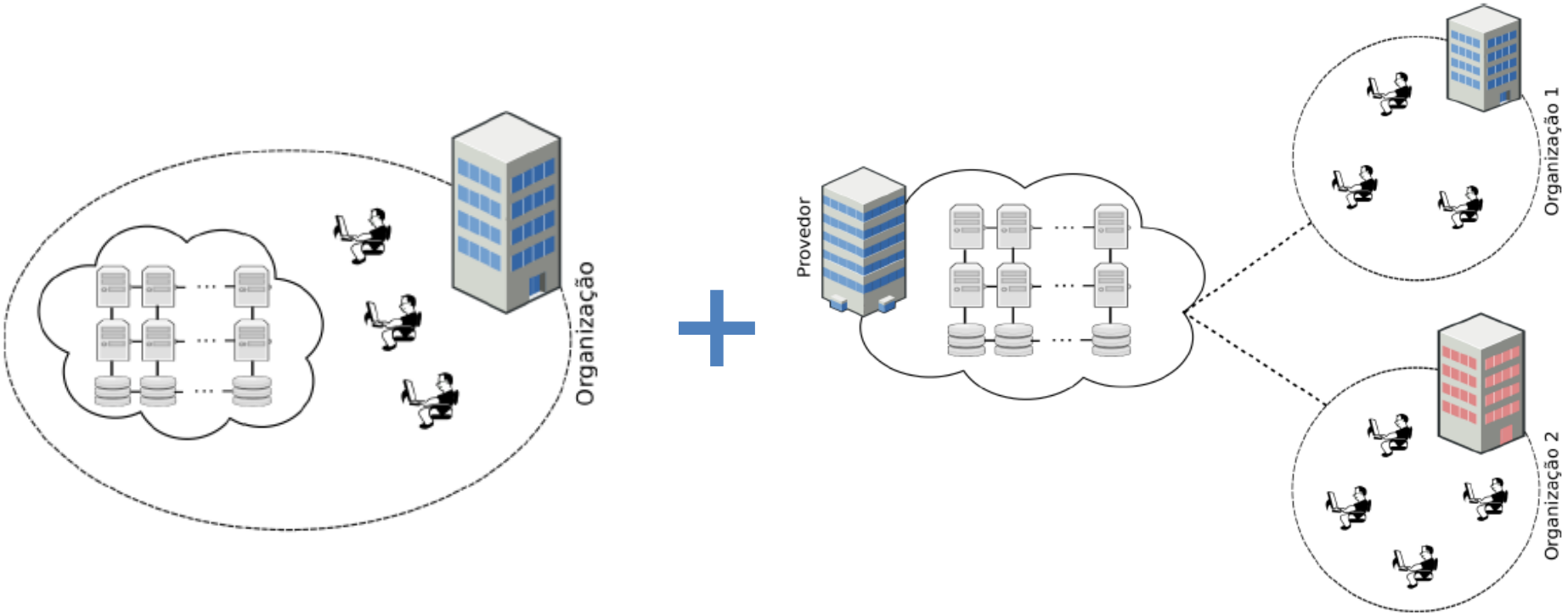
Tipos de Nuvem

- Públicas
- Privadas
- Comunitárias
- Híbridas

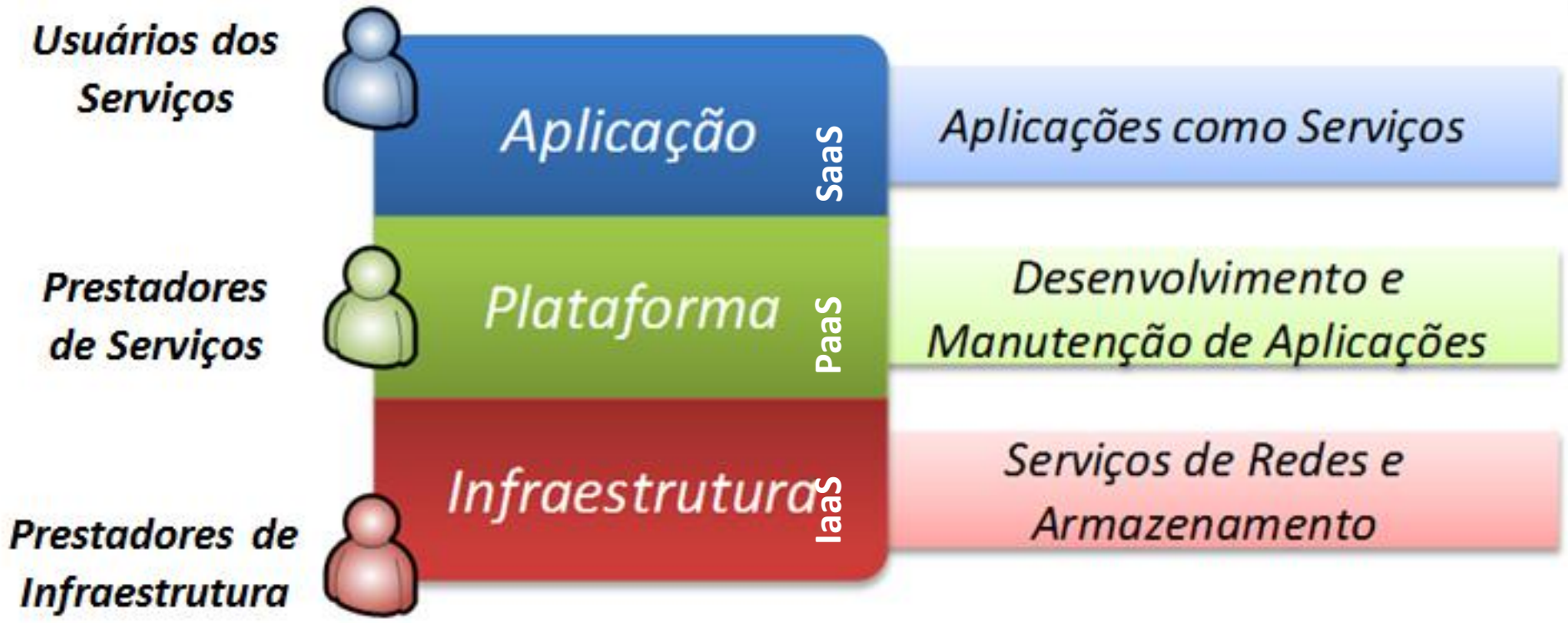


Tipos de Nuvem

- Públicas
- Privadas
- Comunitárias
- Híbridas



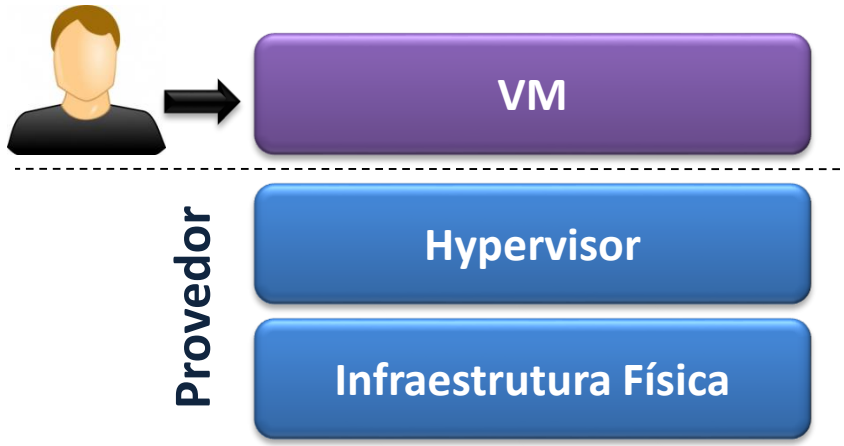
Taxonomia da Nuvem



Fonte: Lamia Youseff, Maria Butrico and Dilma Da Silva"**Toward a Unified Ontology of Cloud Computing**.", Grid Computing Environments Workshop, 2008. GCE '08

IaaS – *Infrastructure as a Service*

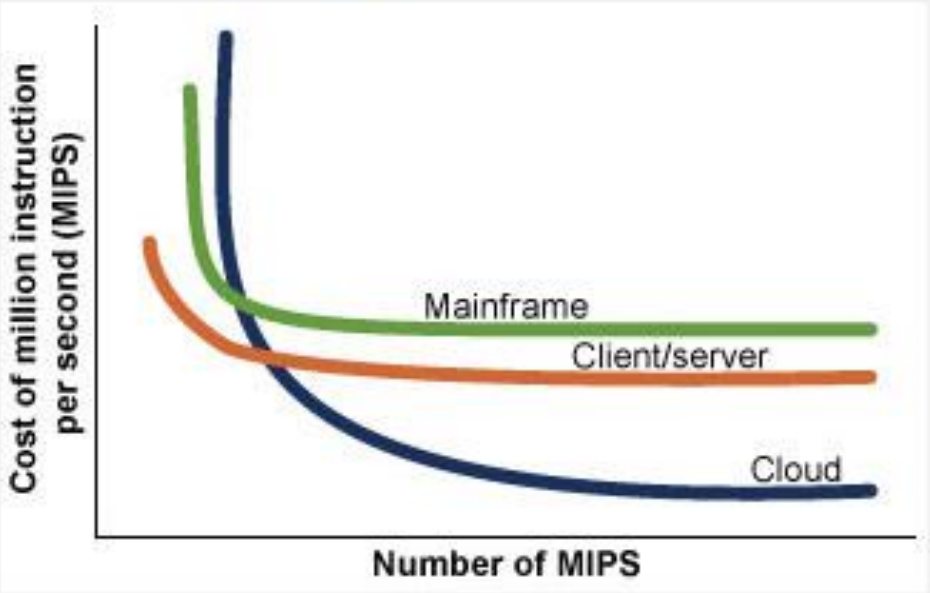
- Oferece **infraestrutura de hardware**
  - Normalmente por meio de virtualização
- Funciona como um aluguel de recursos:
  - Equipamentos de Rede
  - Memória
  - CPU
  - Armazenamento
- Infraestrutura deve ser escalável
  - Aumentar ou diminuir recursos de acordo com a necessidade do cliente



IaaS – Custo Financeiro

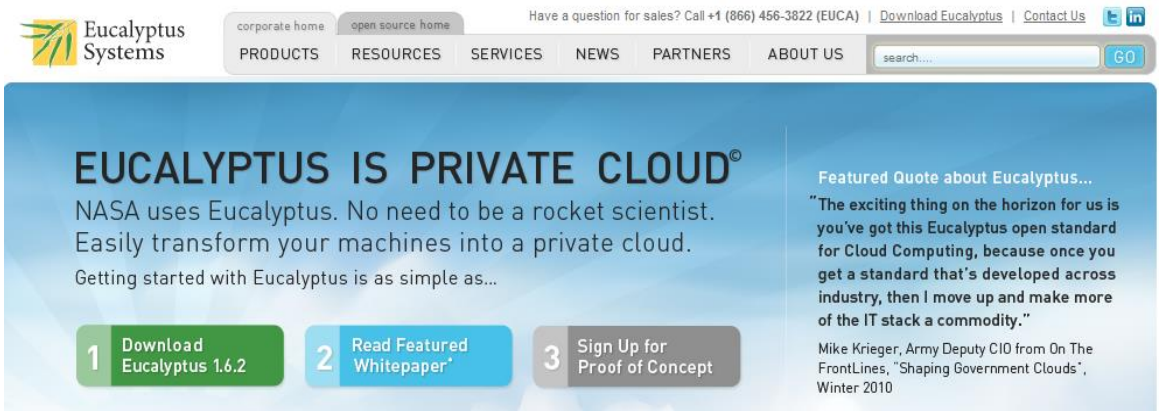
Moving to the cloud can potentially save companies money

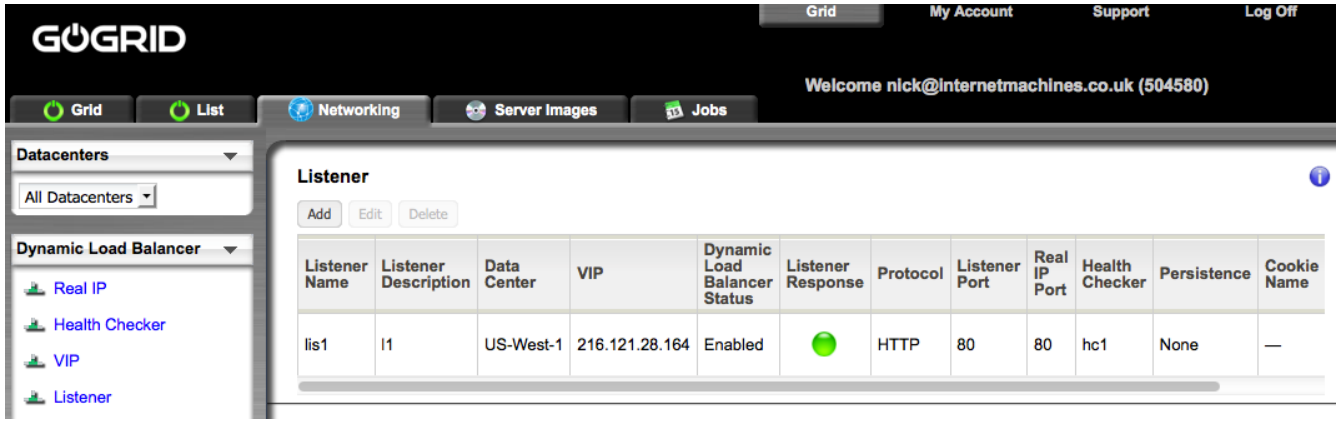
Consolidating computing power has the potential to lead to more efficient computing, and to drive down the cost of processing.



Source: Microsoft.

IaaS: Exemplos





Amazon Elastic Compute Cloud (Amazon EC2)

Amazon EC2

Instâncias on demand padrão	
Pequena (padrão)	$0.060 por hora
Médio	$0.120 por hora
Grande	$0.240 por hora
Extragrande	$0.480 por hora
Instâncias on demand padrão de segunda geração	
Extragrande	$0.500 por hora
Dupla extragrande	$1.000 por hora
Microinstâncias on demand	
Micro	$0.020 por hora
Instâncias on demand com mais memória	
Extragrande	$0.410 por hora
Dupla extragrande	$0.820 por hora
Quádrupla extragrande	$1.640 por hora
Instâncias on demand com CPU de alta performance	
Médio	$0.145 por hora
Extragrande	$0.580 por hora
Instâncias de computação em cluster	
Quádrupla extragrande	$1.300 por hora
Óctupla extragrande	$2.400 por hora
Instâncias on demand de cluster com mais memória	
Óctupla extragrande	$3.500 por hora

Amazon EC2

Instâncias on demand padrão	
Pequena (padrão)	$0.060 por hora
Médio	$0.120 por hora
Grande	$0.240 por hora
Extragrande	$0.480 por hora
Instâncias on demand padrão de se ão	
Extragrande	$0.500 por hora
Dupla extragrande	$1.000 por hora
Microinstâncias on demand	
Micro	$0.020 por hora
Instâncias on demand com mais memória	
Extragrande	$0.410 por hora
Dupla extragrande	$0.820 por hora
Quádrupla extragrande	$1.640 por hora
Instâncias on demand com CPU de alta performance	
Médio	$0.145 por hora
Extragrande	$0.580 por hora
Instâncias de computação em cluster	
Quádrupla extragrande	$1.300 por hora
Óctupla extragrande	$2.400 por hora
Instâncias on demand de cluster com mais memória	
Óctupla extragrande	$3.500 por hora

Instâncias de uso geral

Amazon EC2

Instâncias on demand padrão	
Pequena (padrão)	$0.060 por hora
Médio	$0.120 por hora
Grande	$0.240 por hora
Extragrande	$0.480 por hora
Instâncias on demand padrão de segunda geração	
Extragrande	$0.500 por hora
Dupla extragrande	$1.000 por hora
Microinstâncias on demand	
Micro	$0.020 por hora
Instâncias on demand com mais memória	
Extragrande	$0.410 por hora
Dupla extragrande	$0.820 por hora
Quádrupla extragrande	$1.640 por hora
Instâncias on demand com CPU de alta performance	
Médio	$0.145 por hora
Extragrande	$0.580 por hora
Instâncias de computação em cluster	
Quádrupla extragrande	$1.300 por hora
Óctupla extragrande	$2.400 por hora
Instâncias on demand de cluster com mais memória	
Óctupla extragrande	$3.500 por hora

Amazon EC2

Instâncias on demand padrão	
Pequena (padrão)	$0.060 por hora
Médio	$0.120 por hora
Grande	$0.240 por hora
Extragrande	$0.480 por hora
Instâncias on demand padrão de segunda geração	
Extragrande	$0.500 por hora
Dupla extragrande	$1.000 por hora
Microinstâncias on demand	
Micro	$0.020 por hora
Instâncias on demand com mais memória	
Extragrande	$0.410 p
Dupla extragrande	$0
Quádrupla extragrande	hora
Instâncias on demand com CPU de alta performance	
Médio	$0.145 por hora
Extragrande	$0.580 por hora
Instâncias de computação em cluster	
Quádrupla extragrande	$1.300 por hora
Óctupla extragrande	$2.400 por hora
Instâncias on demand de cl ais memória	
Óctupla extragrande	$3.500 por hora

Instâncias de computação de alto desempenho

Amazon EC2

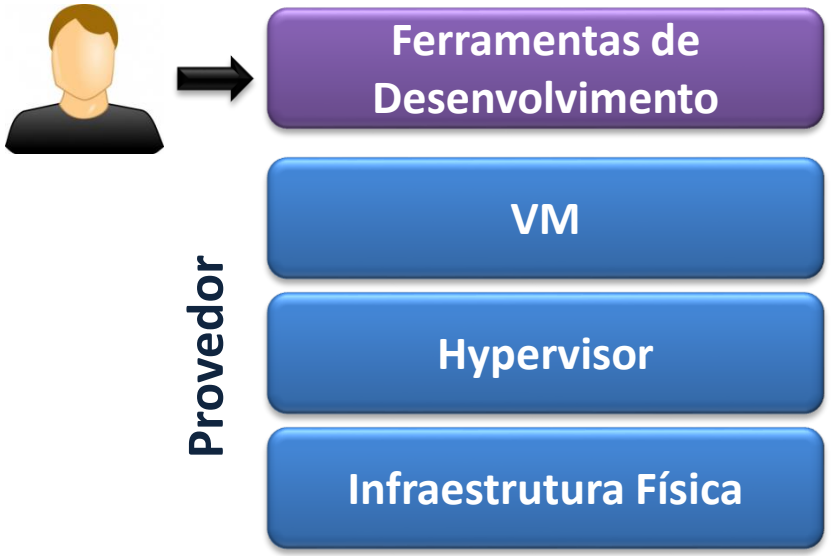
Free Tier*

As part of AWS's Free Usage Tier, new AWS customers can get started with Amazon EC2 for free. Upon sign-up, new AWS customers receive the following EC2 services each month for one year:

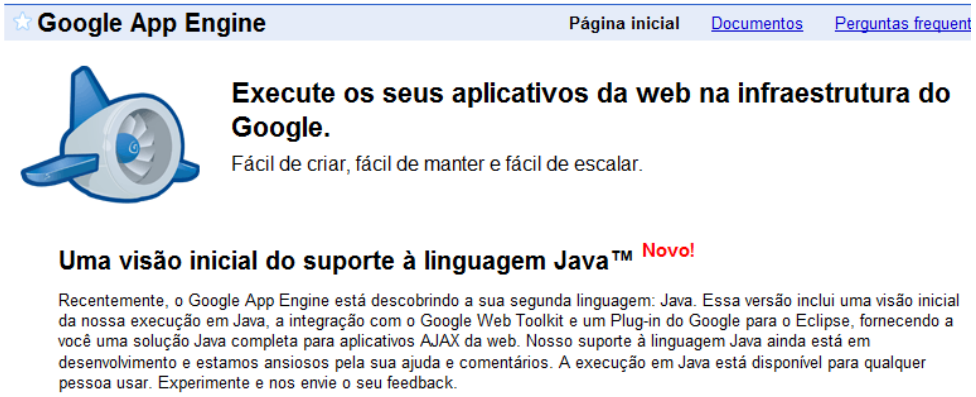
- 750 hours of EC2 running Linux/Unix Micro instance usage
- 750 hours of Elastic Load Balancing plus 15 GB data processing
- 10 GB of Amazon Elastic Block Storage (EBS) plus 1 million IOs, 1 GB snapshot storage, 10,000 snapshot Get Requests and 1,000 snapshot Put Requests
- 15 GB of bandwidth out aggregated across all AWS services
- 1 GB of Regional Data Transfer

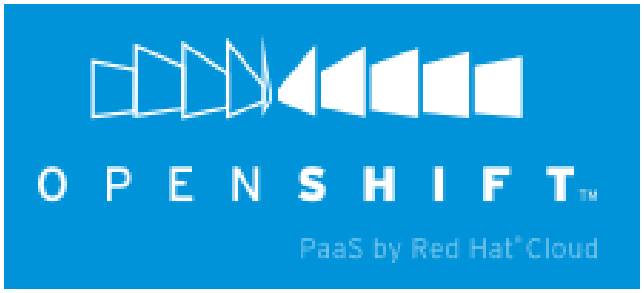
PaaS – Platform as a Service

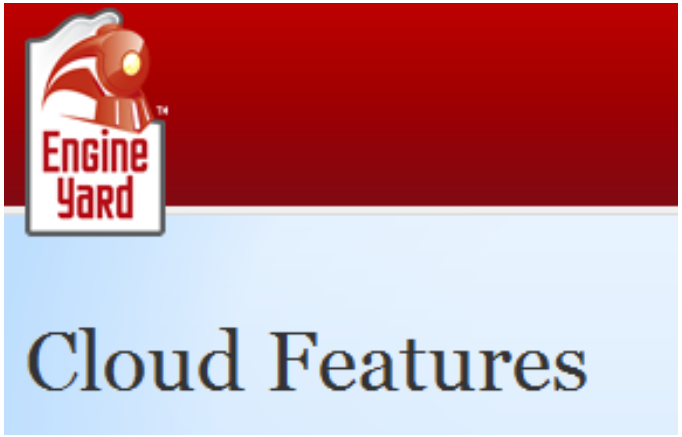
- Modelo onde se fornece **recursos para a construção de aplicações** e serviços para a Internet
- Os recursos incluem:
  - Ferramentas de desenvolvimento
  - Teste
  - Hospedagem
  - Banco de Dados
  - Segurança
  - Controle de versões



PaaS: Exemplos

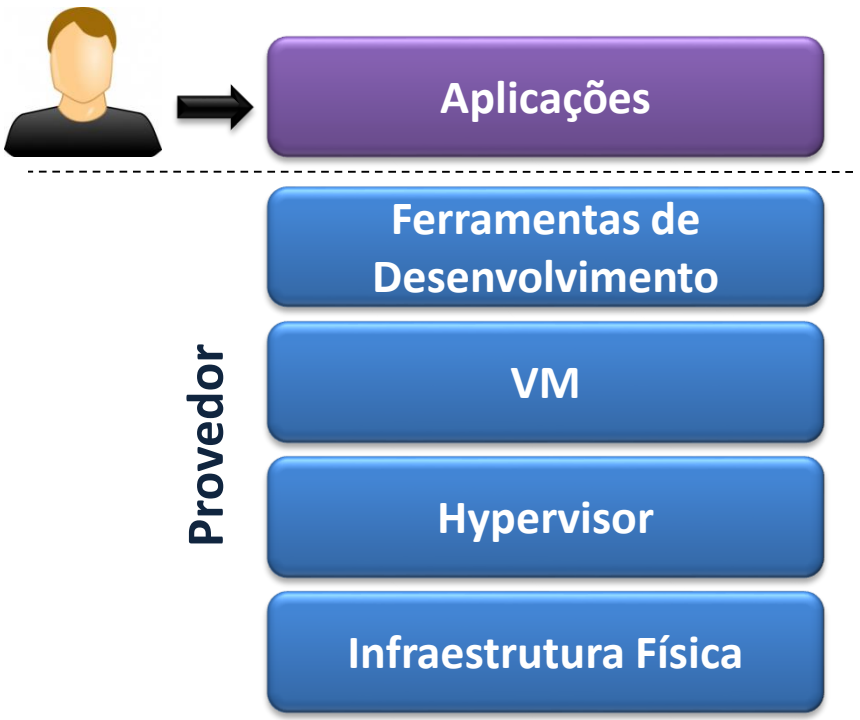






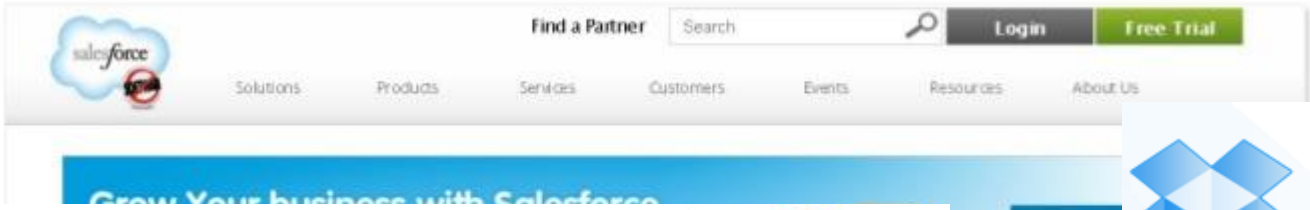
SaaS – Software as a Service

- Modelo no qual uma **aplicação** é armazenada em um servidor
- Usuários a acessam via Internet
  - Não há a necessidade de dar suporte à aplicação

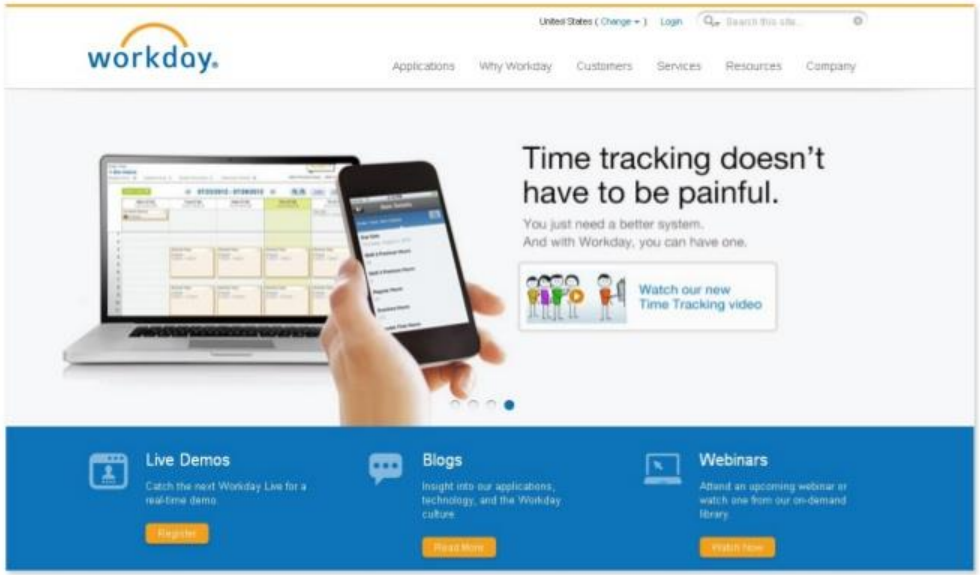


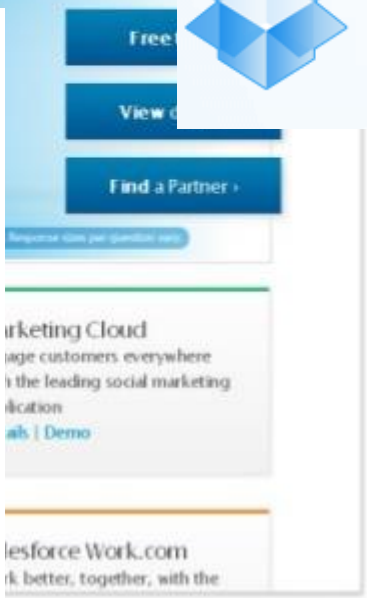
SaaS: Exemplos

Salesforce is a leader is SaaS computing – and is best known for its on-demand Customer Relationship Management (CRM) solutions.



Workday is a leading SaaS provider of human capital management applications, which includes things like payroll and employee expenses management, time tracking and procurement.





Dropbox

Google Docs

p://cloudcomputingwire.com

Casos de Sucesso

- The New York Times
  - Amazon EC2 e S3 (100 instâncias)
  - Conversão de 11 milhões de artigos (4TB)
-  
  - Processamento de vídeos do Big Brother Brasil
  - Evita sobrecarga nos servidores locais

Casos na área científica

- Análise Filogenômica

OLIVEIRA, D., Ocaña, K. A. C. S., Ogasawara, E., Dias, J., Goncalves, J., Mattoso, M., (2012), "**Cloud-based Phylogenomic Inference of Evolutionary Relationships: A Performance Study**". 2nd International Workshop on Cloud Computing and Scientific Applications (CCSA), Ottawa.

SciPhylomics

Phylogenomic analysis

- Estudos Evolutivos

Ocaña, K. A. C. S., OLIVEIRA, D. de, Horta, F., Dias, J., Ogasawara, E., Mattoso, M., (2012), "Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific Workflow". Brazilian Symposium on Bioinformatics (BSB 2012), Campo Grande, MT.

SciEvol

Evolutionary analysis

- Genômica Comparativa

OCANA, K. ; OLIVEIRA, D. ; DIAS, J. ; OGASAWARA, E. ; MATTOSO, M. L. Q. . Optimizing Phylogenetic Analysis Using SciHmm Cloud-based Scientific Workflow. 7th IEEE e Science conference. IEEE Computer Society, 2011.

SciHmm

Comparative genomics

Roteiro do Tutorial

- Motivação
- Workflows Científicos
- Nuvens de Computadores
- **Proveniência de Dados**
- Máquinas de Workflow para Nuvem
- Aplicação de Proveniência em e-Science
- Demo

O que é proveniência?

Dictionary

prov•e•nance |ˈprävənəns|

noun

the place of origin or earliest known history of something : *an orange rug of Iranian provenance.*

- the beginning of something's existence; something's origin : *they try to understand the whole universe, its provenance and fate.*

See note at **ORIGIN** .

- a record of ownership of a work of art or an antique, used as a guide to authenticity or quality : *the manuscript has a distinguished provenance.*

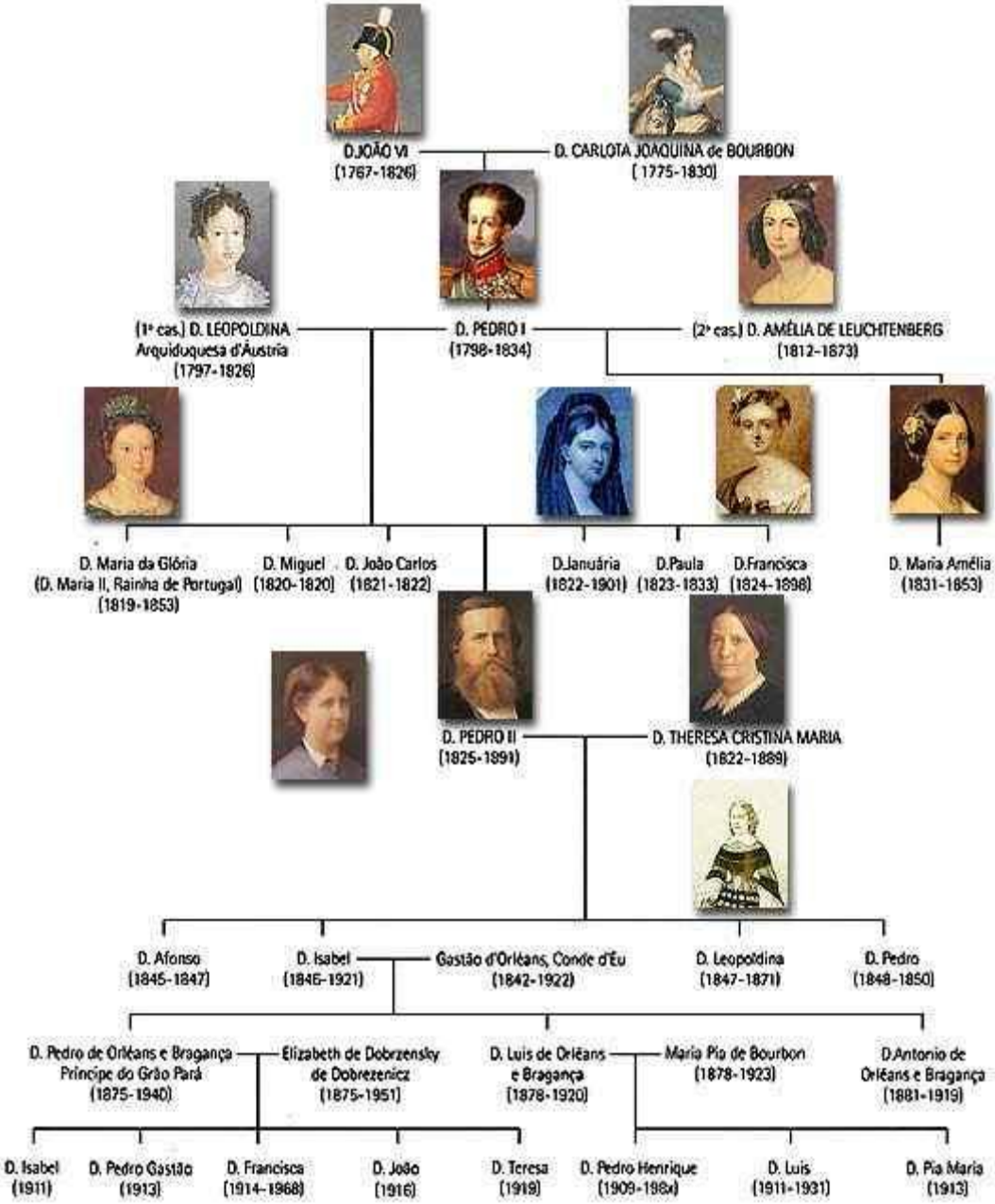
ORIGIN late 18th cent.: from French, from the verb ***provenir*** *'come or stem from,'* from Latin ***provenire***, from ***pro-*** *'forth'* + ***venire*** *'come.'*

O que é proveniência?

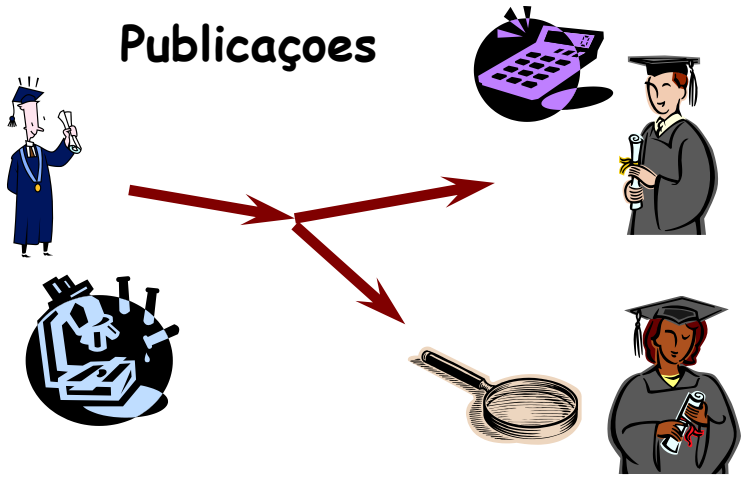
- De acordo com o dicionário Aurélio:
  - *"s.f. Primeira manifestação; começo, princípio. Procedência; ponto de partida de uma nação, de uma família; naturalidade. Nascimento, proveniência; ascendência. Origem."*

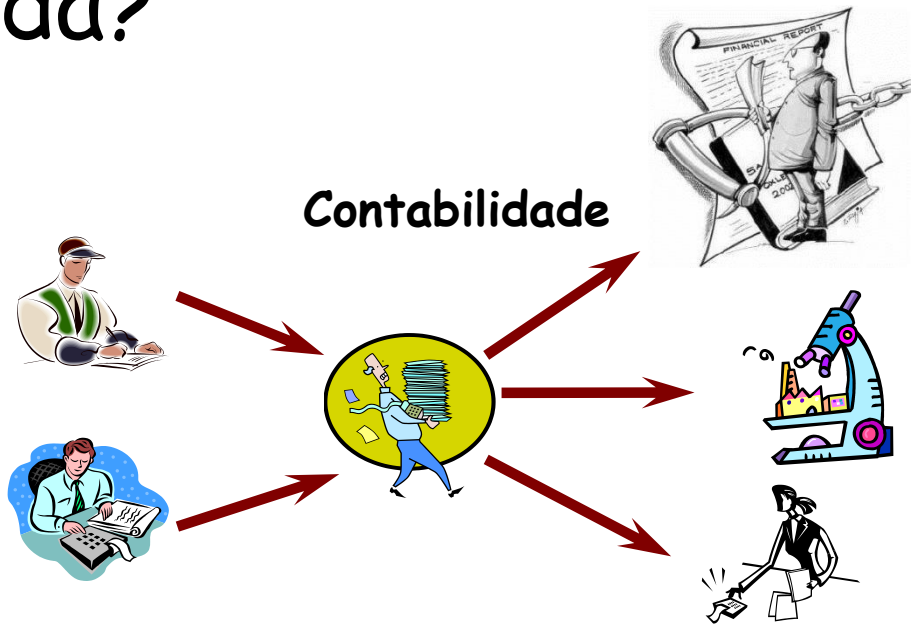
C ?

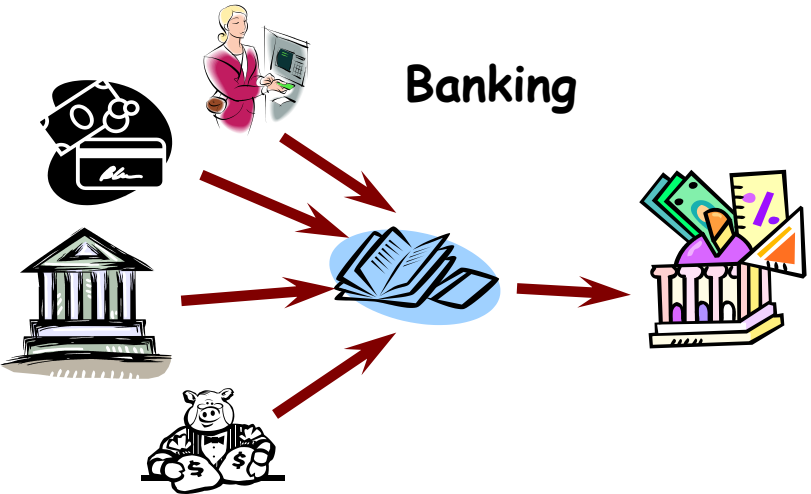
- De acc io:
  - *"s.f.), prin rtida de uma dade. Nasc ncia. Orig*

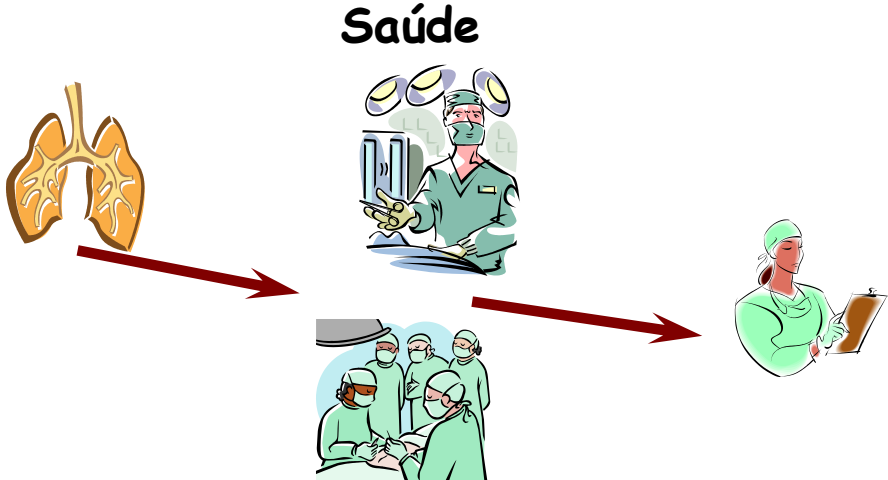


Onde a proveniência pode ser aplicada?









Proveniência nas artes

ARTE E CULTURA

09 de Setembro de 2013 • 10h22

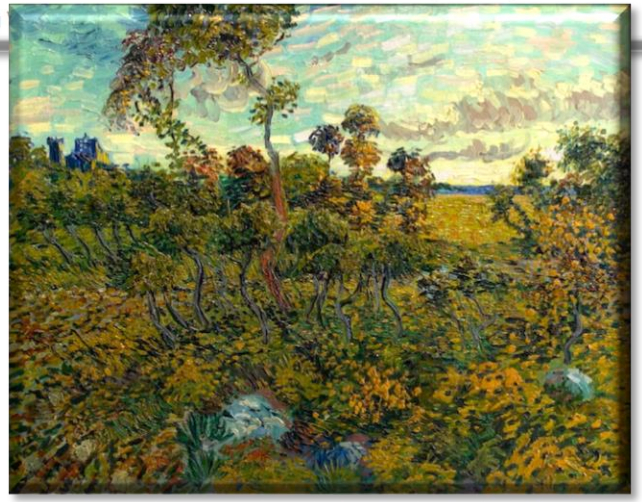
Após um século escondido, quadro "falso" é confirmado como Van Gogh

3 Tweetar 4 Curtir

Uma paisagem francesa que passou um século guardada em um sótão por ser considerada uma tela falsa de Vincent van Gogh, na verdade, é um trabalho autêntico, segundo o resultado de um novo estudo divulgado nesta segunda-feira.

A tela "Ocaso em Montmajour", que mostra azinheiras retorcidas e uma distante ruína banhada pela luz do entardecer, foi pintada em 1888, quando Van Gogh vivia em Arles, no sul da França.

A pintura, pertencente a um colecionador privado, será exposta a partir deste mês, durante um ano, no Museu



- Pesquisa a autoria de obras de arte em acervos de museus ou coleções particulares
- Dados de proveniência podem incluir restaurações realizadas nas obras
- Ajudam especialistas a verificar autenticidade de obras

Adaptação dos slides de Freire e Davidson – SIGMOD 2008

Proveniência nas artes

- Pesquisa a autoria de obras de arte em acervos de museus ou coleções particulares
- Dados de proveniência podem incluir restaurações realizadas nas obras
- Ajudam especialistas a verificar autenticidade de obras

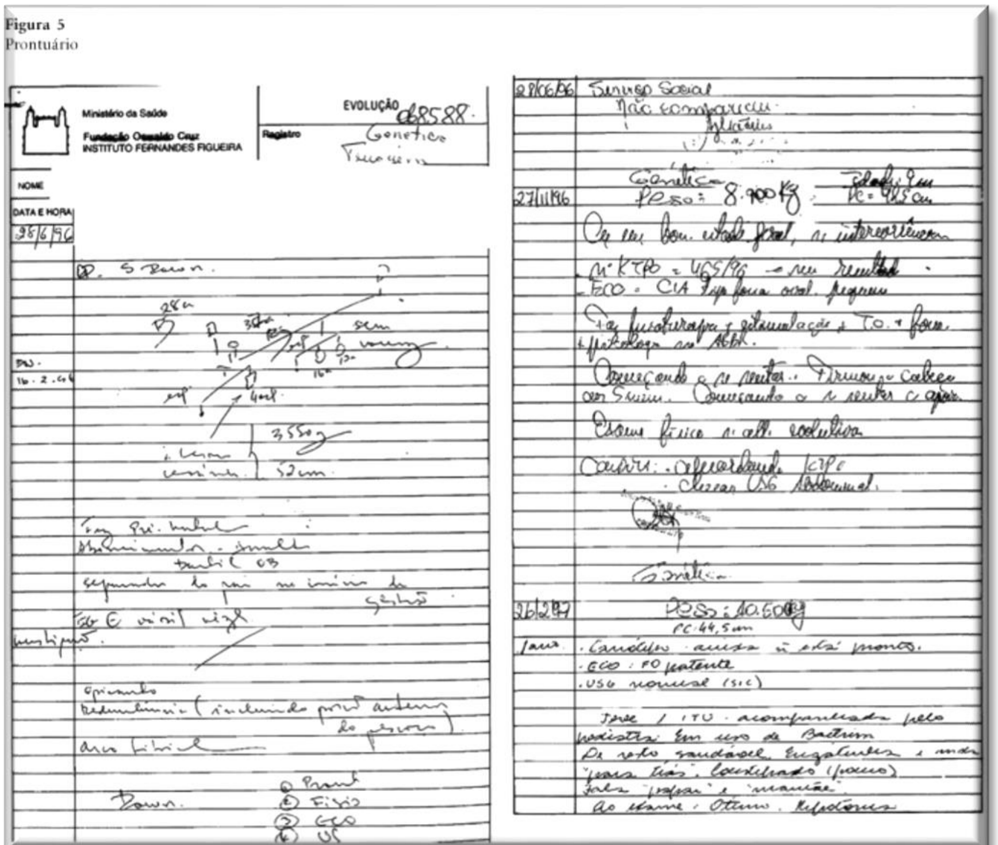
Adaptação dos slides de Freire e Davidson – SIGMOD 2008

Proveniência na indústria

- Garante a procedência do alimento
- É possível saber em que fazenda, granja etc o alimento foi produzido ou colhido

Adaptação dos slides de Freire e Davidson – SIGMOD 2008

Proveniência na medicina



- Garante o rastreio das mudanças do paciente
- Cirurgias realizadas, procedimentos realizados, medicamentos prescritos

fonte: A epistemologia narrativa e o exercício clínico do diagnóstico, Maria Helena Cabral de Almeida Cardoso, Kenneth Rochel de Camargo Jr., Juan Clinton Llerena Jr. , Ciência & Saúde Coletiva

Adaptação dos slides de Freire e Davidson – SIGMOD 2008

Proveniência na Ciência

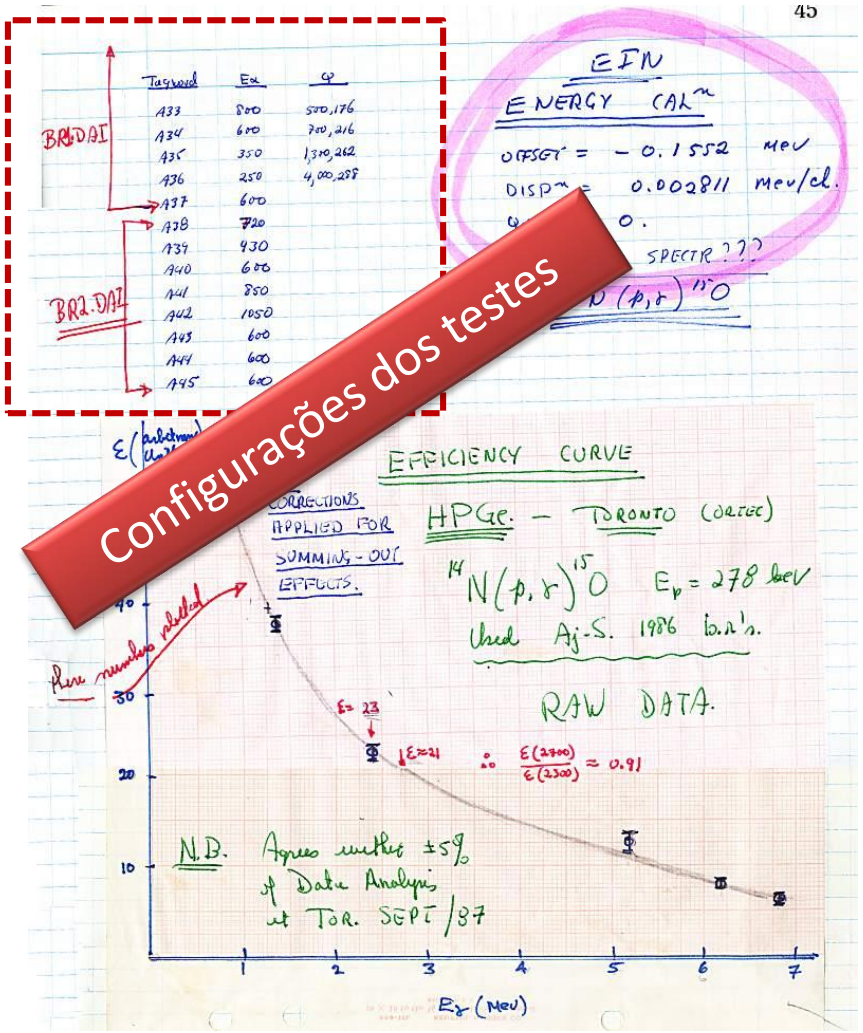
- Reprodução e interpretação de um resultado
- Compreensão do raciocínio empregado no experimento
- Verificação se o experimento segui procedimentos específicos
- Rastreio de responsabilidades
  - Quem fez o que e quando?

Proveniência na Ciência

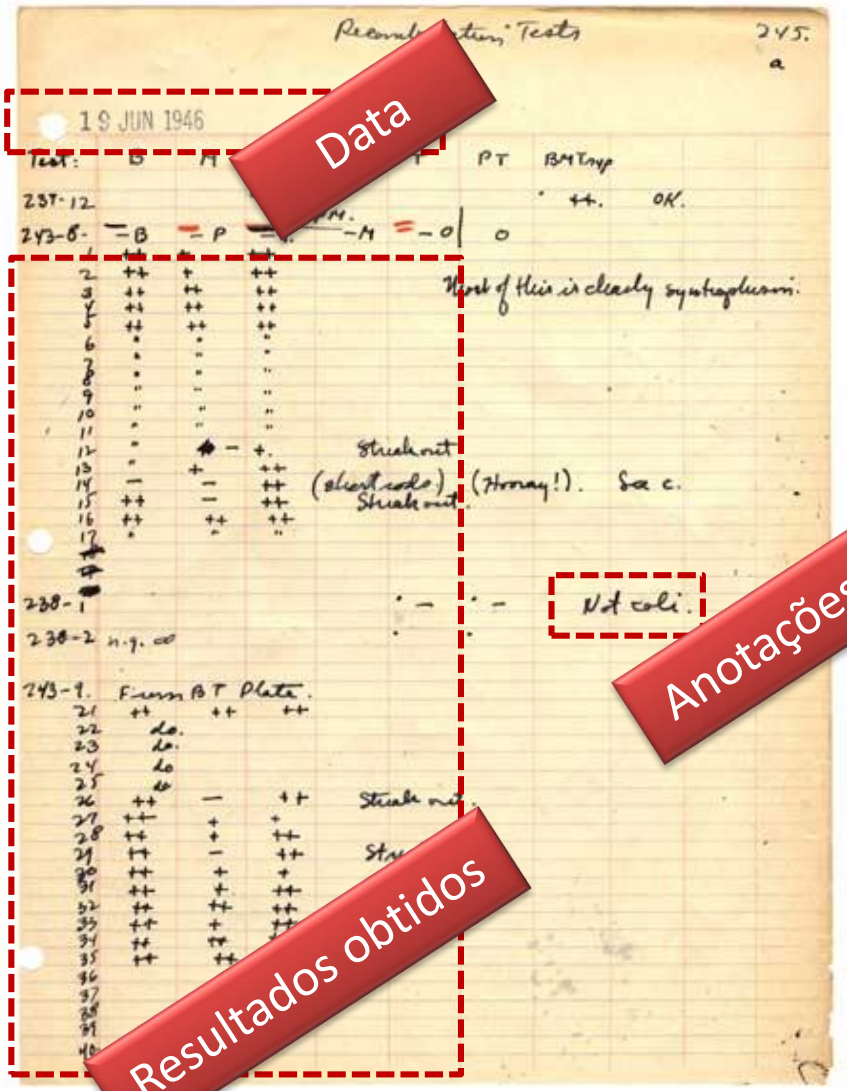
- Reprodução e interpretação de u resultado
- Compreensão do raci do no experimento
- Verificaç mento segui pro pecíficos
- responsabilidades
  - m fez o que e quando?

"Provenance is as (or more) important as the results"
(Juliana Freire e Susan Davidson, SIGMOD 2008)

Proveniência na Ciência



Configurações dos testes



Data

Anotações

Resultados obtidos

Por que a curadoria de dados científicos é importante em e-Science?

- Faz parte do processo de pesquisa
- Agrega valores *Intrínsecos* e *Extrínsecos* à pesquisa
- Aumenta o potencial de criar "novos conhecimentos" a partir de dados pré-existentes
- Preserva o contexto da descoberta científica

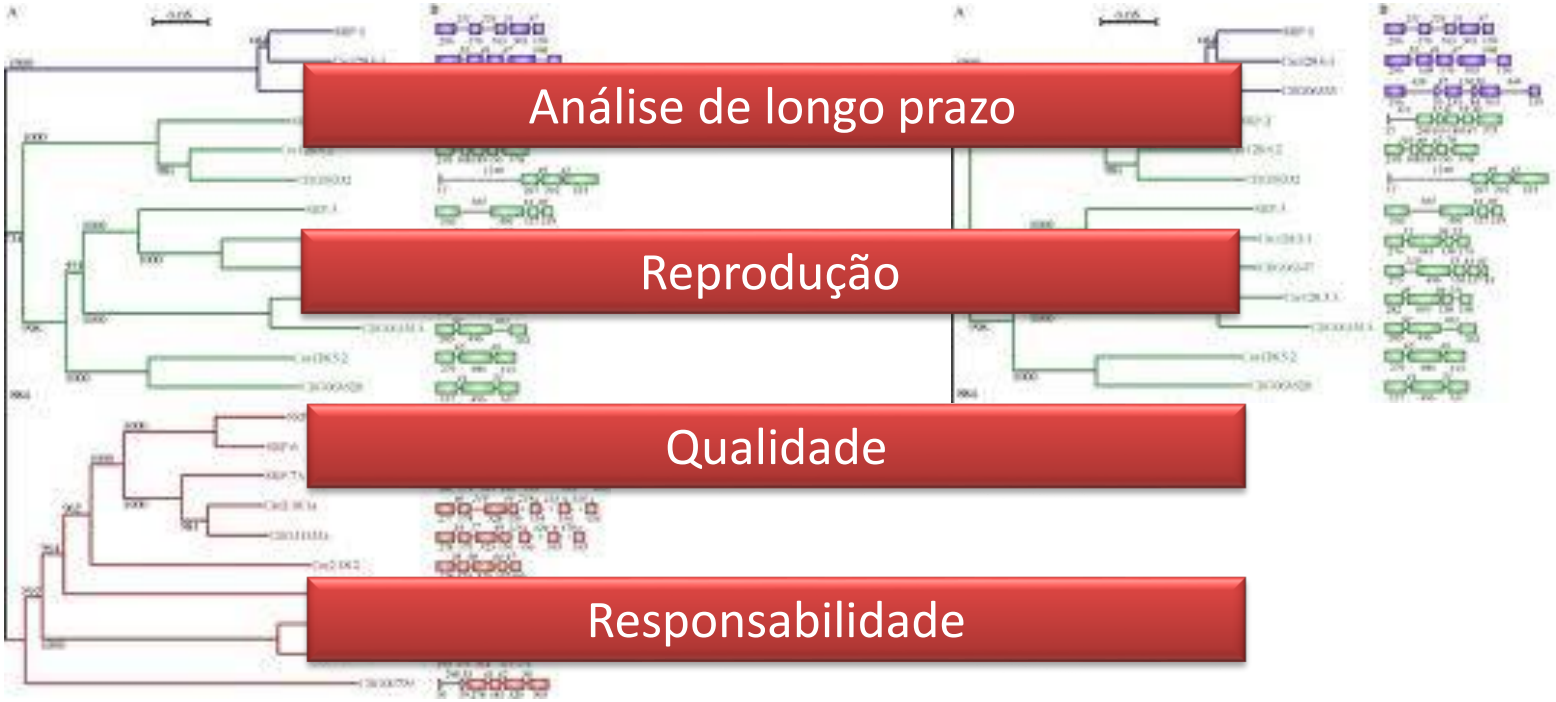
Qualidade dos dados

- A proveniência dos dados permite avaliar a qualidade deles para uma aplicação
- Erros introduzidos por defeitos nos dados tendem a se tornar mais graves quando propagados
- O nível de detalhe da proveniência determina com que grau a qualidade dos dados pode ser estimada

Proveniência em experimentos *in silico*

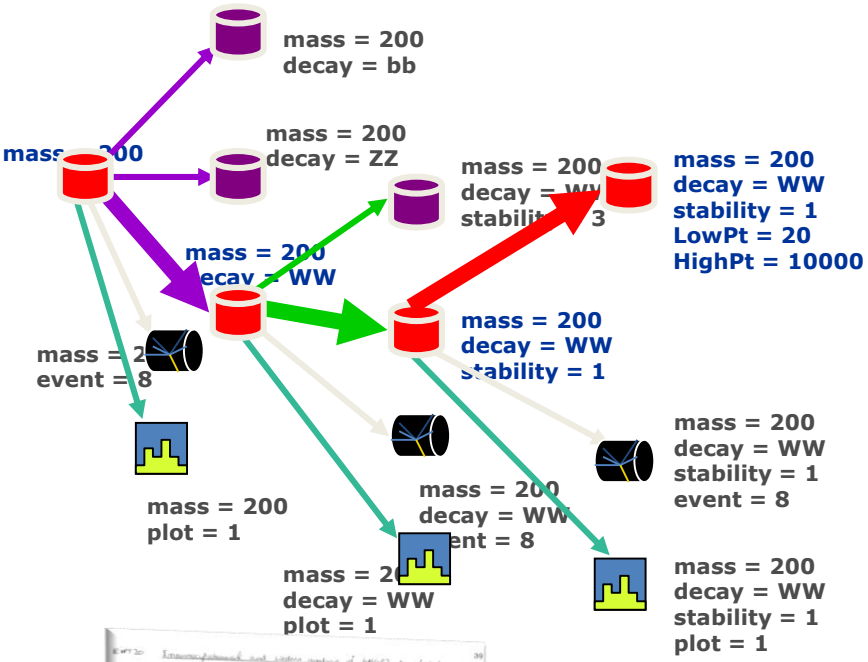
iii-201102281EHWP25EX8.tcoffeeStockholm

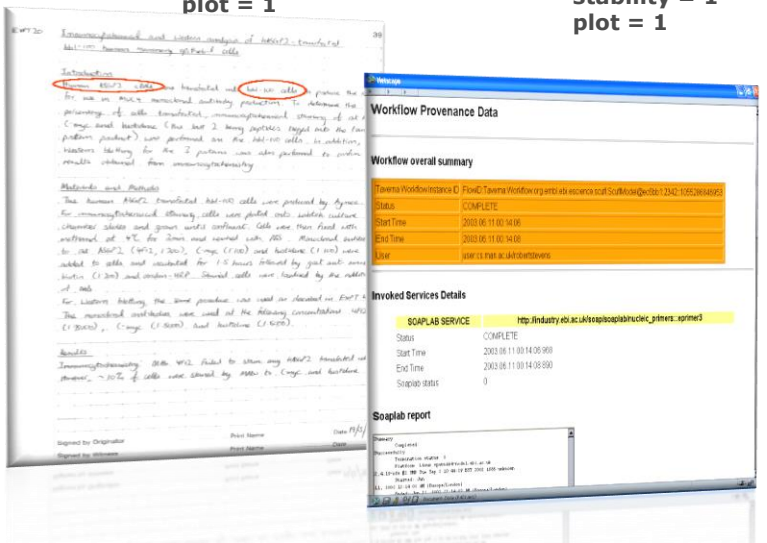
iii-201102281EHWP25EX9.tcoffeeStockholm



Formas da proveniência

- Derivação
  - Representação de um caminho, workflow, roteiro, ou consulta.
  - Associação entre itens (grafos)
  - Geralmente é uma explicação (2W1H =when, who, how)
  - Centrada no processo
- Anotação
  - Anexado aos items ou coleções de dados (estruturado ou semi ou texto livre)
  - Geralmente é uma explicação (5W1H = why, when, where, who, what, how)
  - Centrada no dado





Fonte: Knowledge and Provenance (Goble, 2003)

Proveniência em Tempo Real

- Uso da proveniência em tempo real possibilita análises mais precisas e uma resposta mais rapida durante a execução do experimento
  - As derivações que ocorrem durante a execução do workflow devem estar disponíveis assim que os dados forem produzidos

Dados específicos de domínio também são necessários

Tipos de proveniência ?

A Survey of Data Provenance in e-Science

Yogesh L. Simmhan Beth Plale Dennis Gannon

Computer Science Department
Indiana University, Bloomington, IN 47405
{ysimmhan, plale, gannon}@cs.indiana.edu

ABSTRACT

Provenance of e-Science Experiments - experience from Bioinformatics

Mark Greenwood, Carole Goble, Robert Stevens, Jun Zhao, Matthew Addis, Darren Marvin, Luc Moreau, Tom Oinn
EPSRC e-Science Pilot Project "myGrid"
http://www.mygrid.org.uk

COMPUTATIONAL PROVENANCE

Provenance for Computational Tasks: A Survey

Provenance in Scientific Workflow Systems

Susan Davidson, Sarah Cohen-Boulakia, Anat Eyal
University of Pennsylvania
{susan, sarahcb, anate}@cis.upenn.edu

Bertram Ludäscher, Timothy McPhillips
Shawn Bowers, Manish Kumar Anand
University of California, Davis
{ludaesch, tmcphillips, sbowers, manand}@ucdavis.edu

Juliana Freire
University of Utah
juliana@cs.utah.edu

Abstract

CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE

Tracking Provenance in a

Between Experiments and Data

Briefings in Bioinformatics Advance Access published May 14, 2007

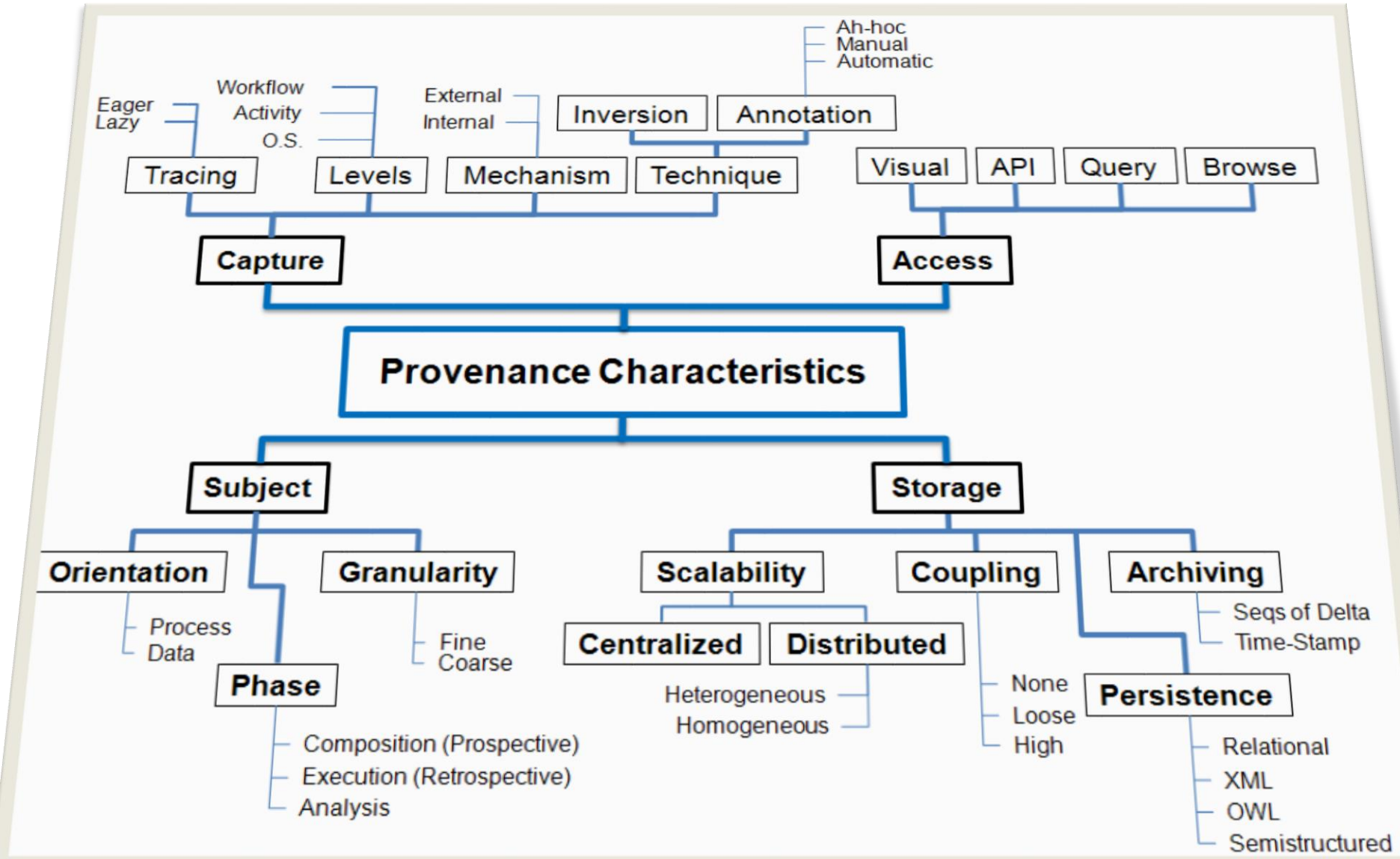
Using provenance to manage knowledge of *In Silico* experiments

Robert Stevens, Jun Zhao and Carole Goble

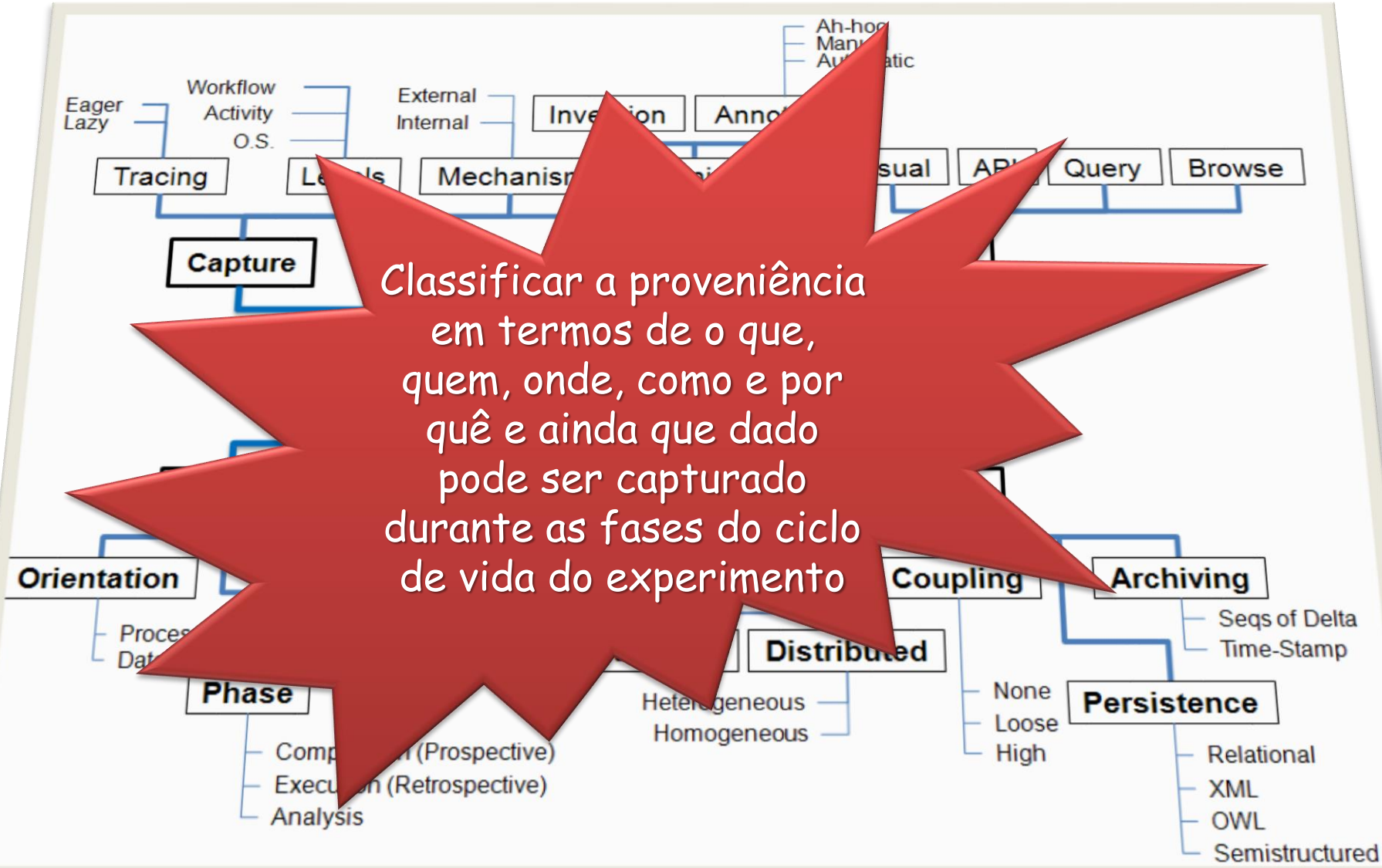
Abstract

This article offers a briefing in one of the knowledge management issues of *in silico* experimentation in bioinformatics. Recording of the provenance of an experiment—what was done; where, how and why, etc. is an important aspect of scientific best practice that should be extended to *in silico* experimentation. We will do this in the context of eScience which has been part of the move of bioinformatics towards an industrial setting. Despite the computational nature of bioinformatics, these analyses are scientific and thus necessitate their own versions of typical scientific rigour. Just as recording who, what, why, when, where and how of an experiment is central to the scientific

Taxonomia de Proveniência

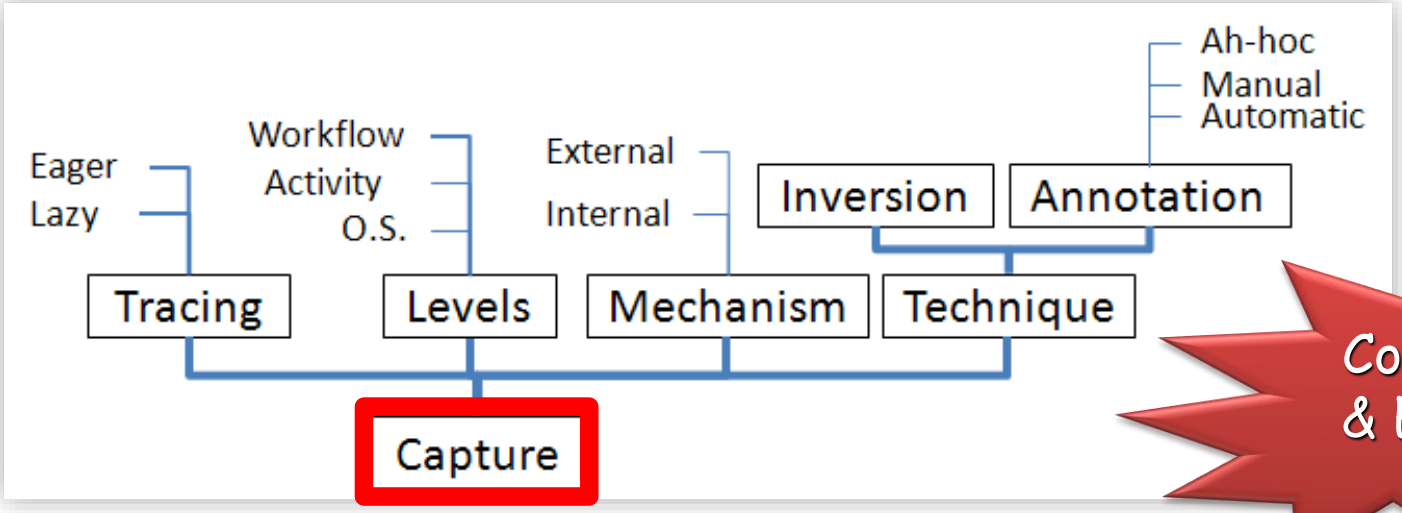


Taxonomia de Proveniência



Taxonomia de Proveniência

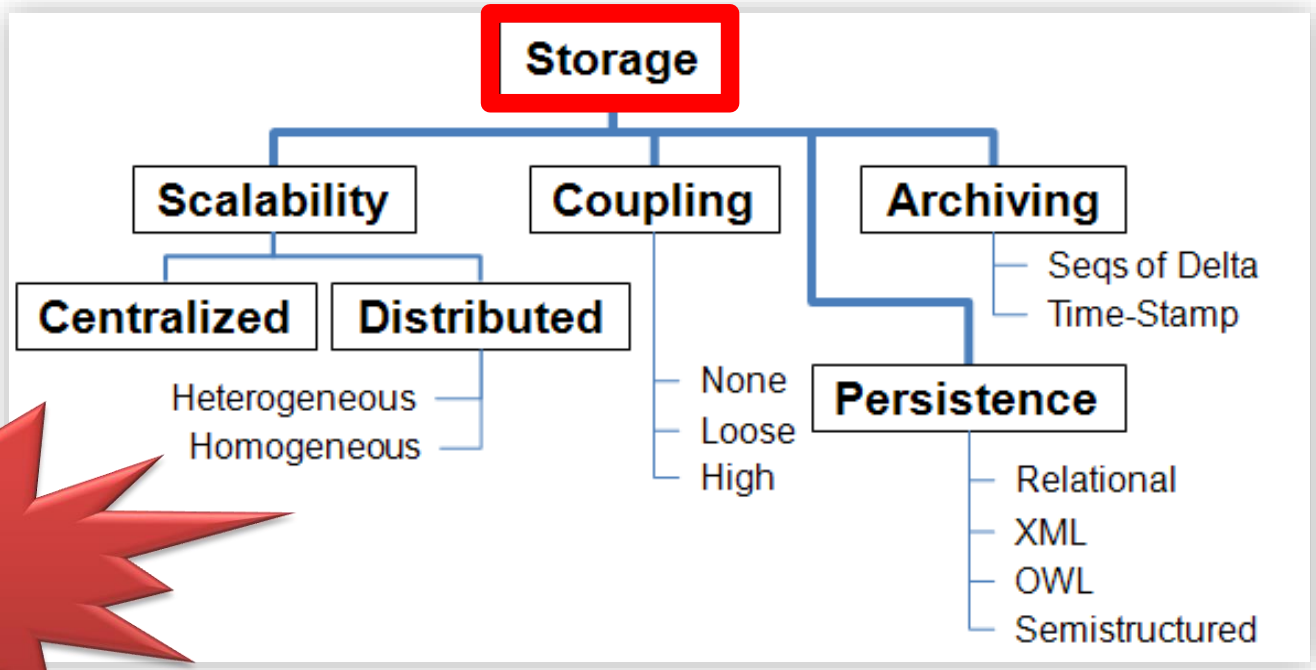
- Classificar como os dados de proveniência podem ser capturados nos sistemas de proveniência.



Cruz, S.M.S. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . Towards a Taxonomy of Provenance in Scientific Workflow Management Systems. In: IEEE SWF, 2009, Los Angeles.

Taxonomia de Proveniência

- Formas de registro e armazenamento dos dados de proveniência.

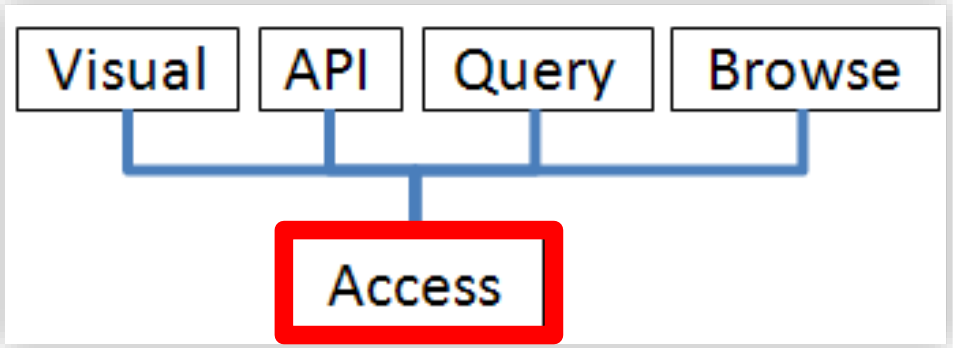


Execução

Cruz, S.M.S. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . Towards a Taxonomy of Provenance in Scientific Workflow Management Systems. In: IEEE SWF, 2009, Los Angeles.

Taxonomia de Proveniência

- Descreve as formas de acesso aos dados e repositórios de proveniência.

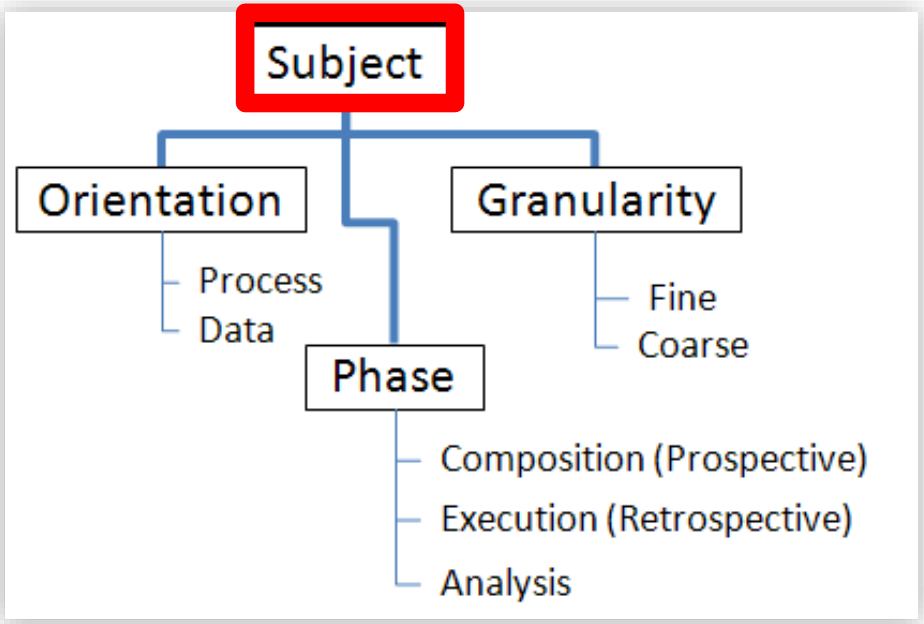


Análise

Cruz, S.M.S. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . Towards a Taxonomy of Provenance in Scientific Workflow Management Systems. In: IEEE SWF, 2009, Los Angeles.

Taxonomia de Proveniência

- Representação da proveniência em termos de assunto e granulosidade.



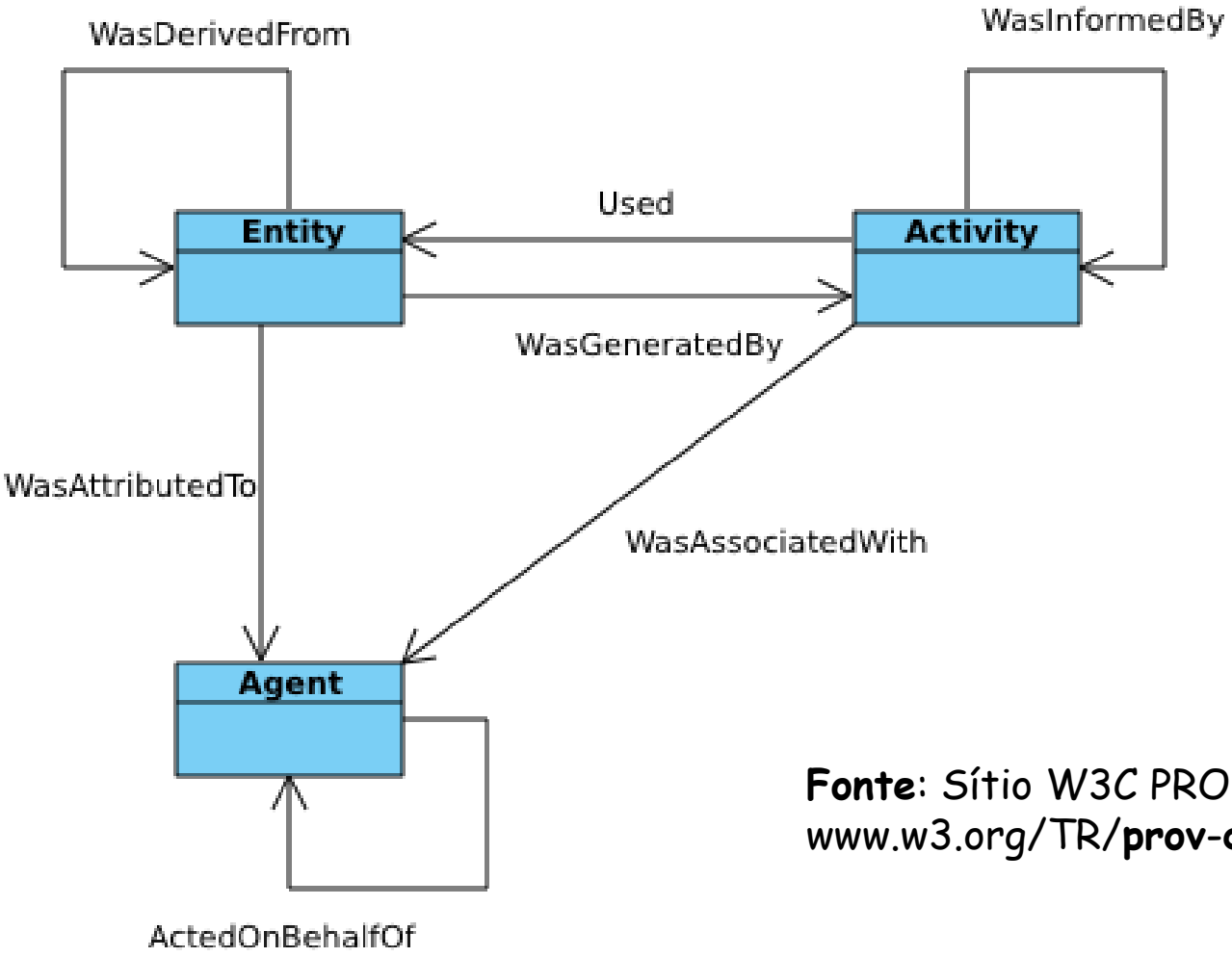
Ortogonal à todas as fases

Cruz, S.M.S. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . Towards a Taxonomy of Provenance in Scientific Workflow Management Systems. In: IEEE SWF, 2009, Los Angeles.

Representação de proveniência com o PROV

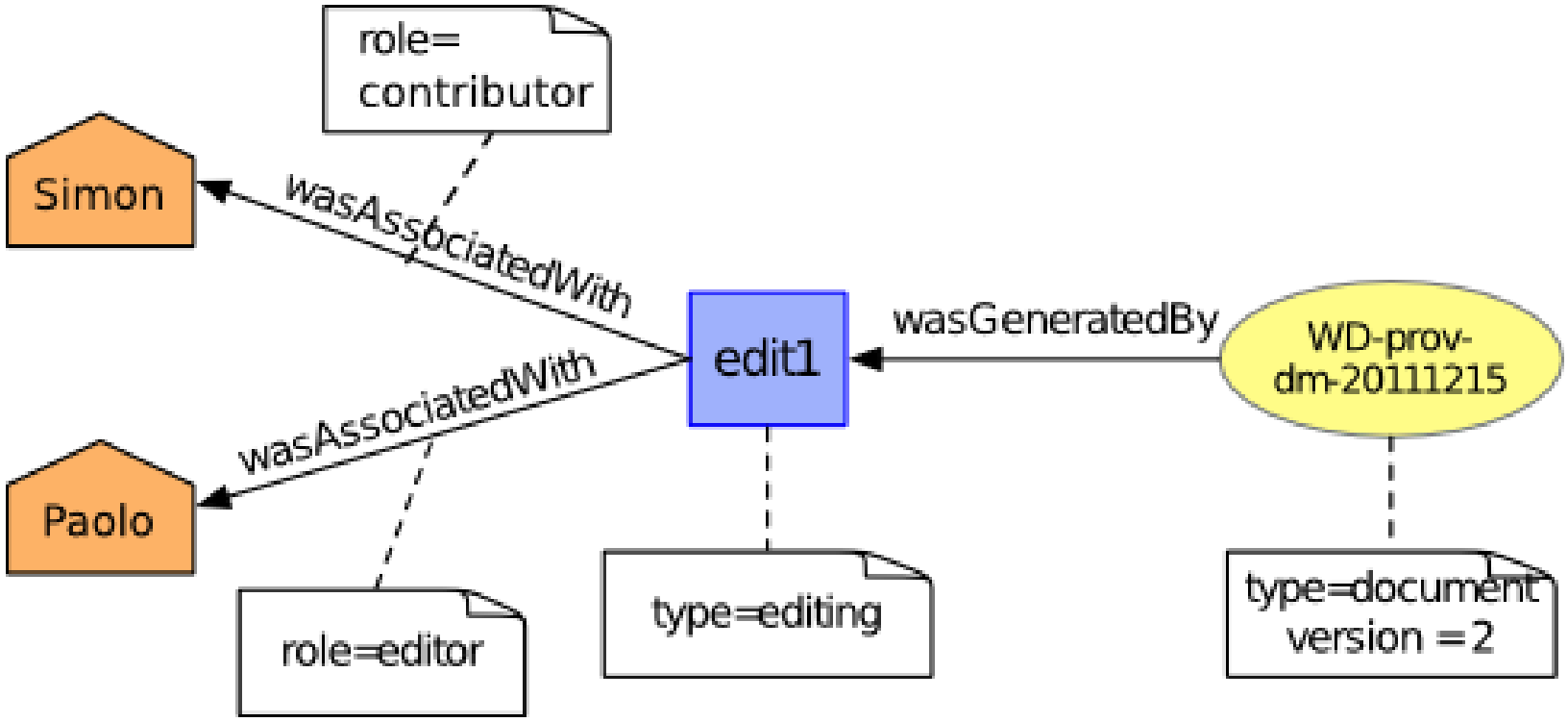
- Padrão W3C para representação de proveniência
- Baseado em 3 elementos principais:
  - Entidade (*Entity*)
  - Agente (*Agent*)
  - Atividade (*Activity*)
- *"aims the inter-operable interchange of provenance information in heterogeneous environments such as the Web"*

O modelo de dados do PROV



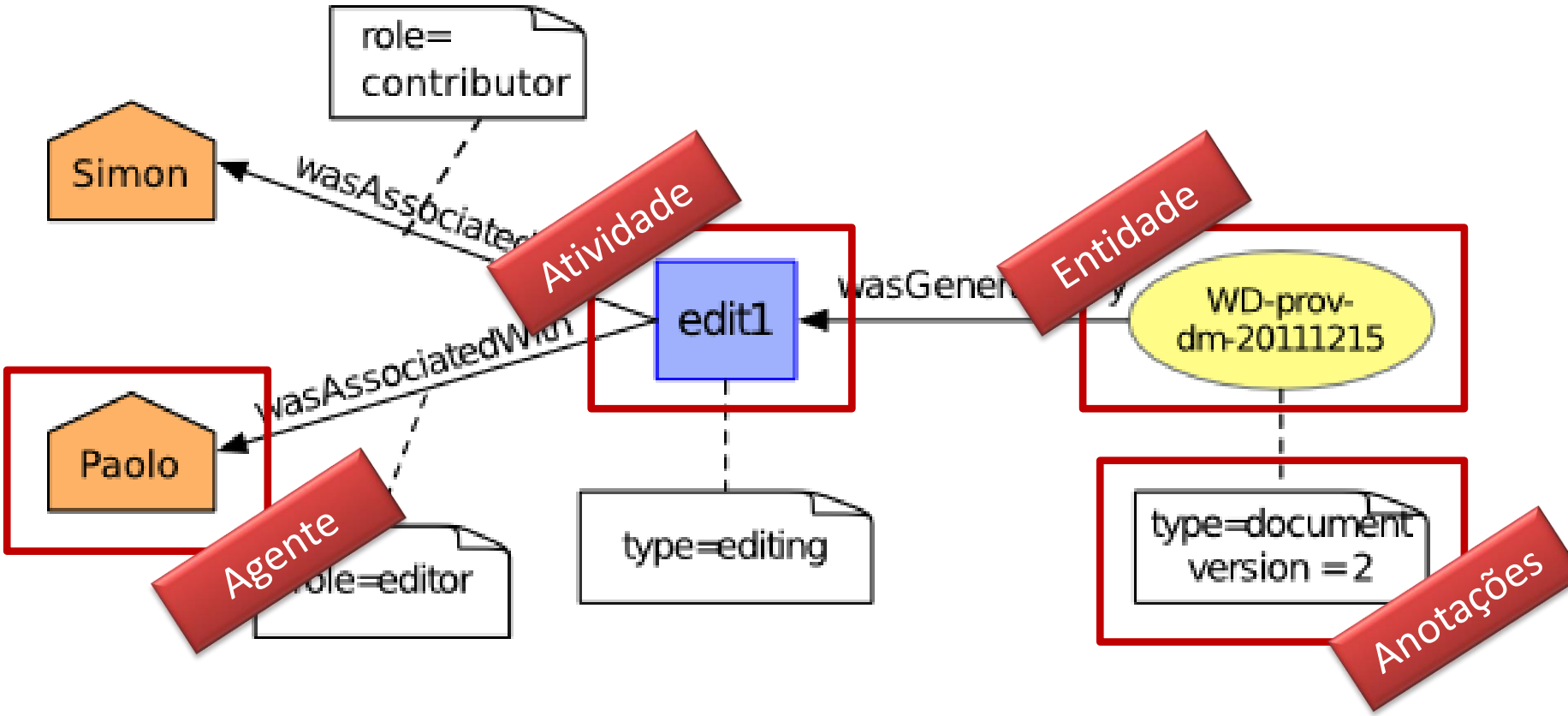
Fonte: Sítio W3C PROV
www.w3.org/TR/**prov-dm**/

Exemplo de utilização do PROV



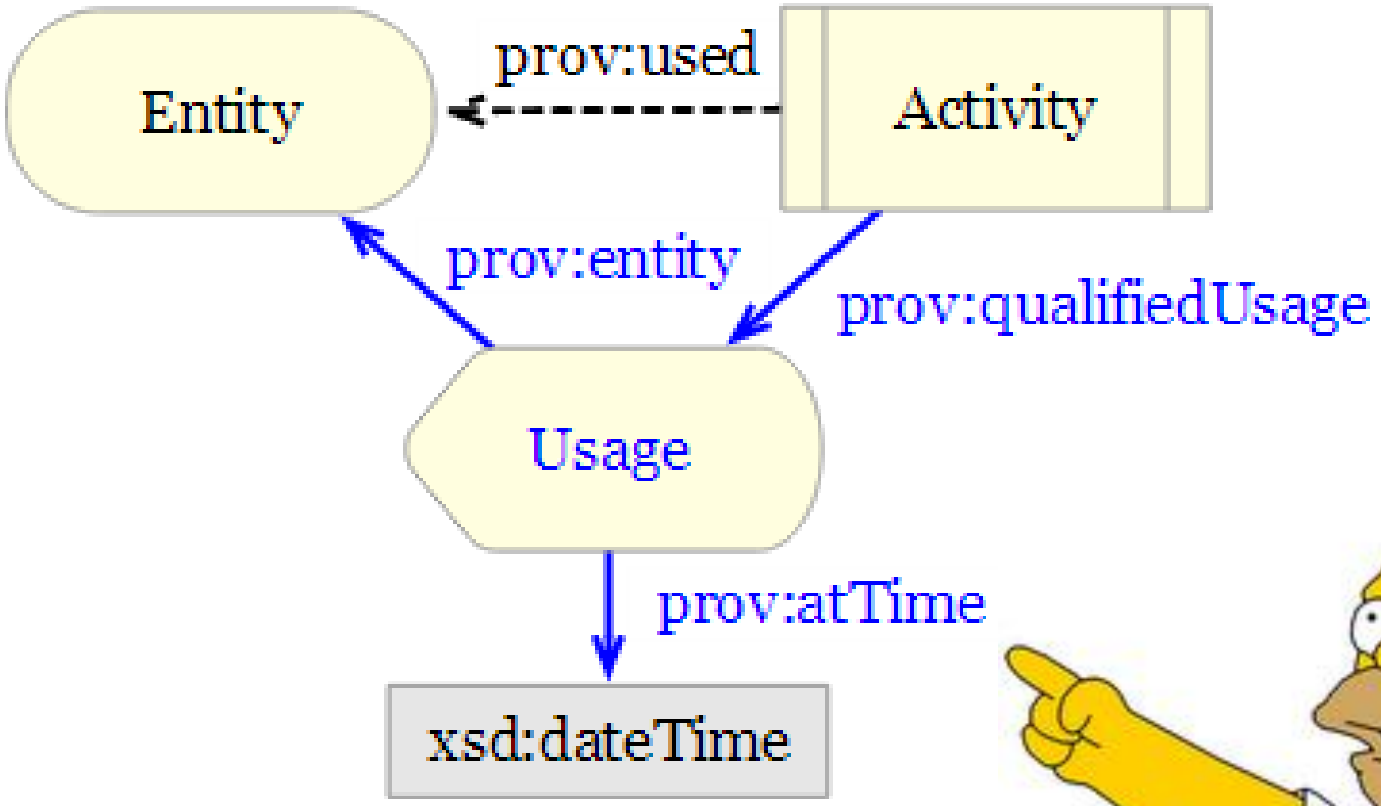
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Exemplo de utilização do PROV



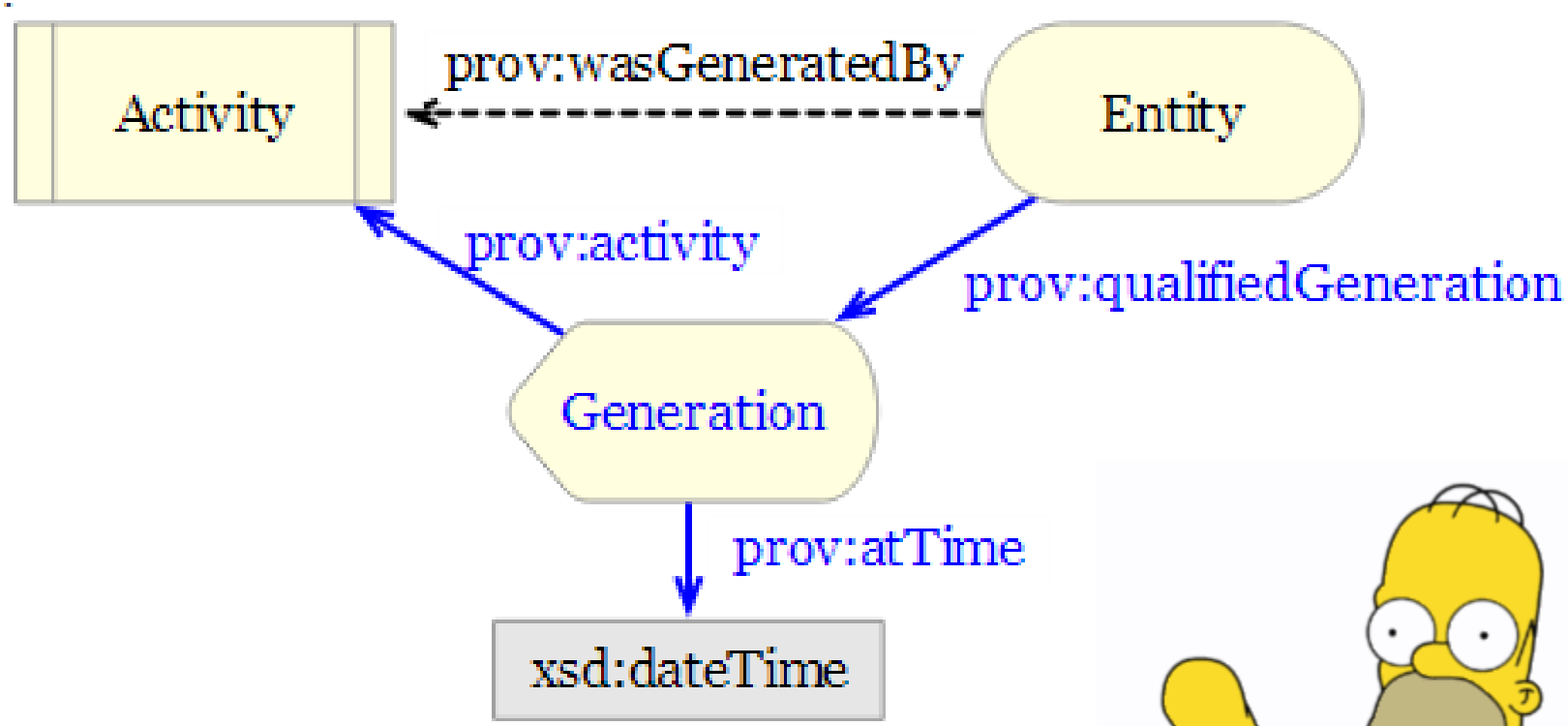
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *used*



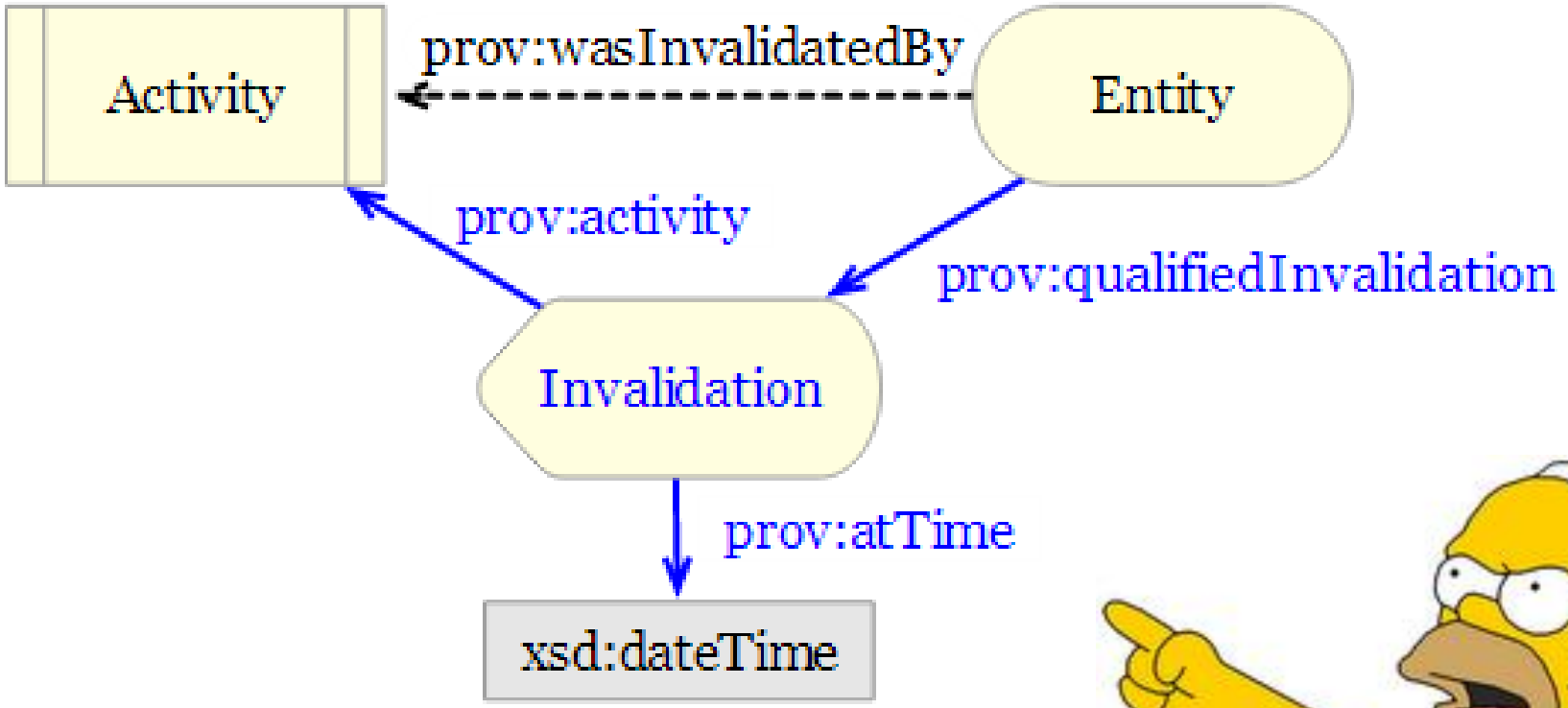
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *wasGeneratedBy*



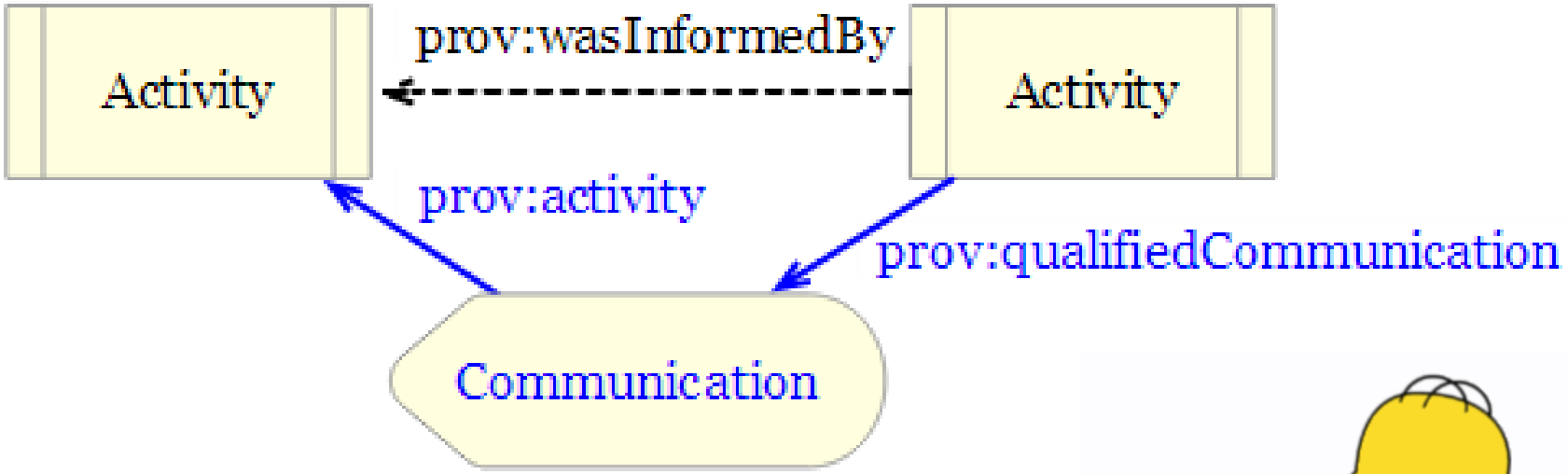
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *wasInvalidatedBy*



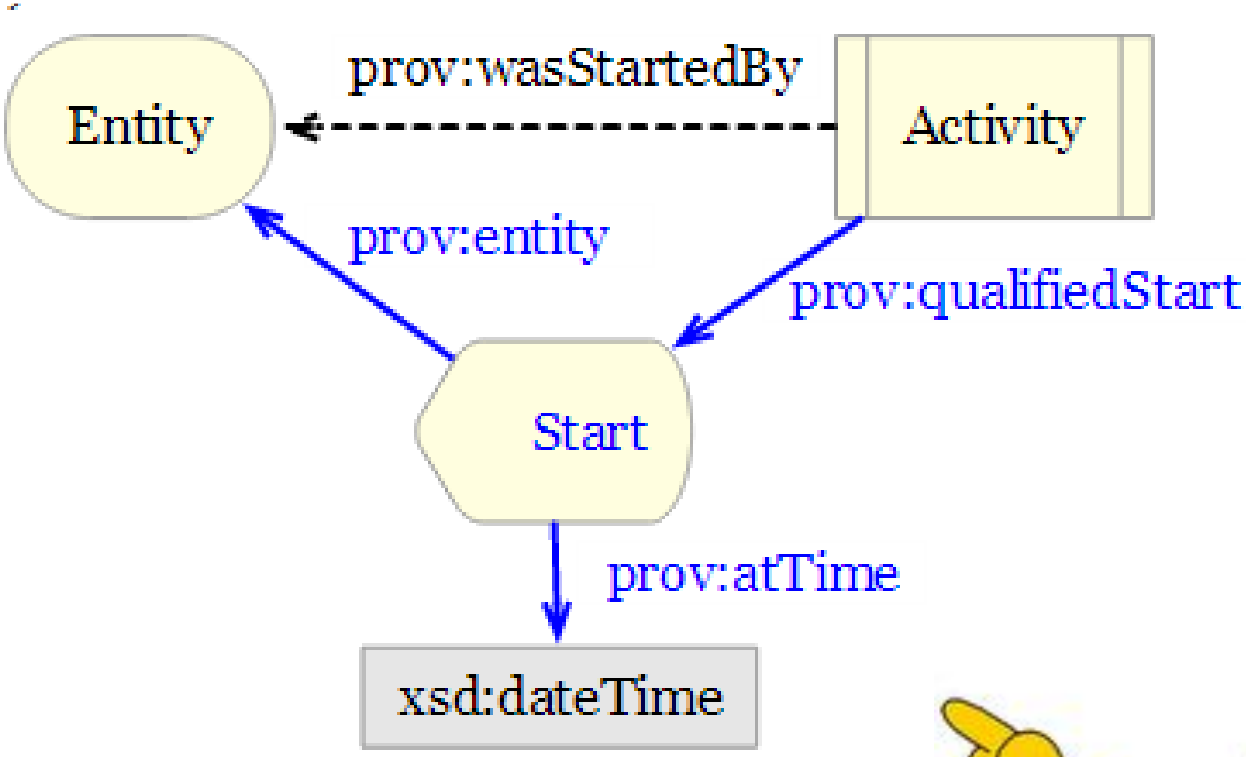
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *wasInformedBy*



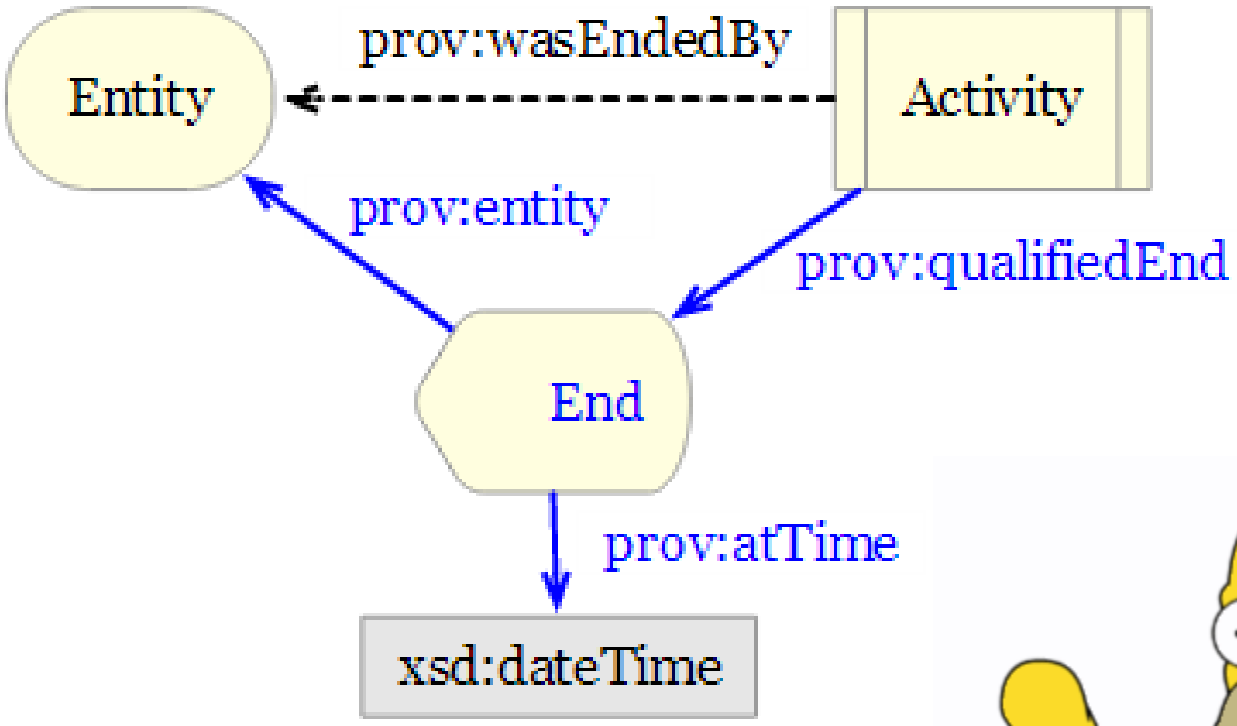
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *wasStartedBy*



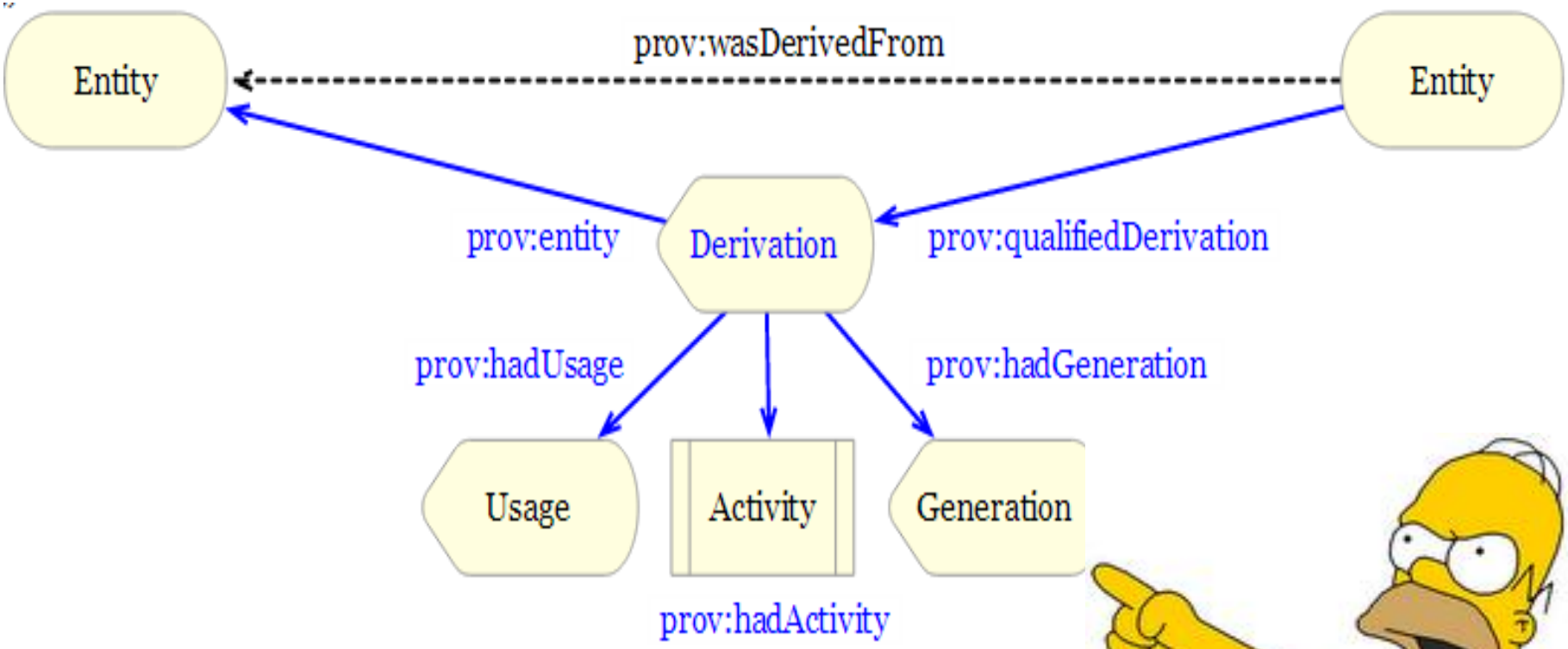
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *wasEndedBy*



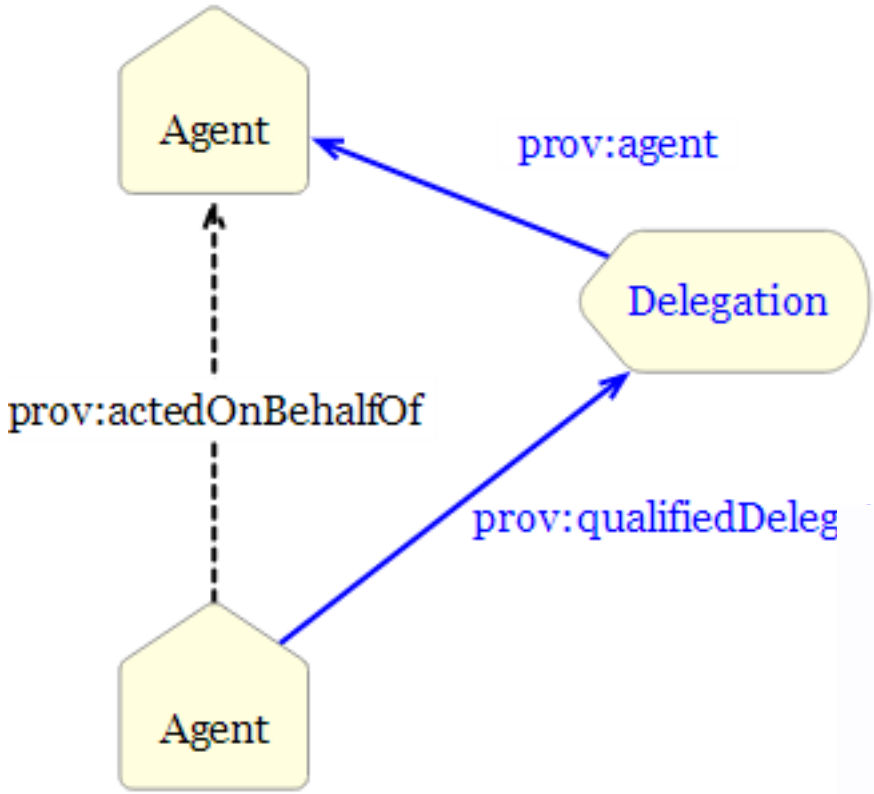
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *wasDerivedFrom*



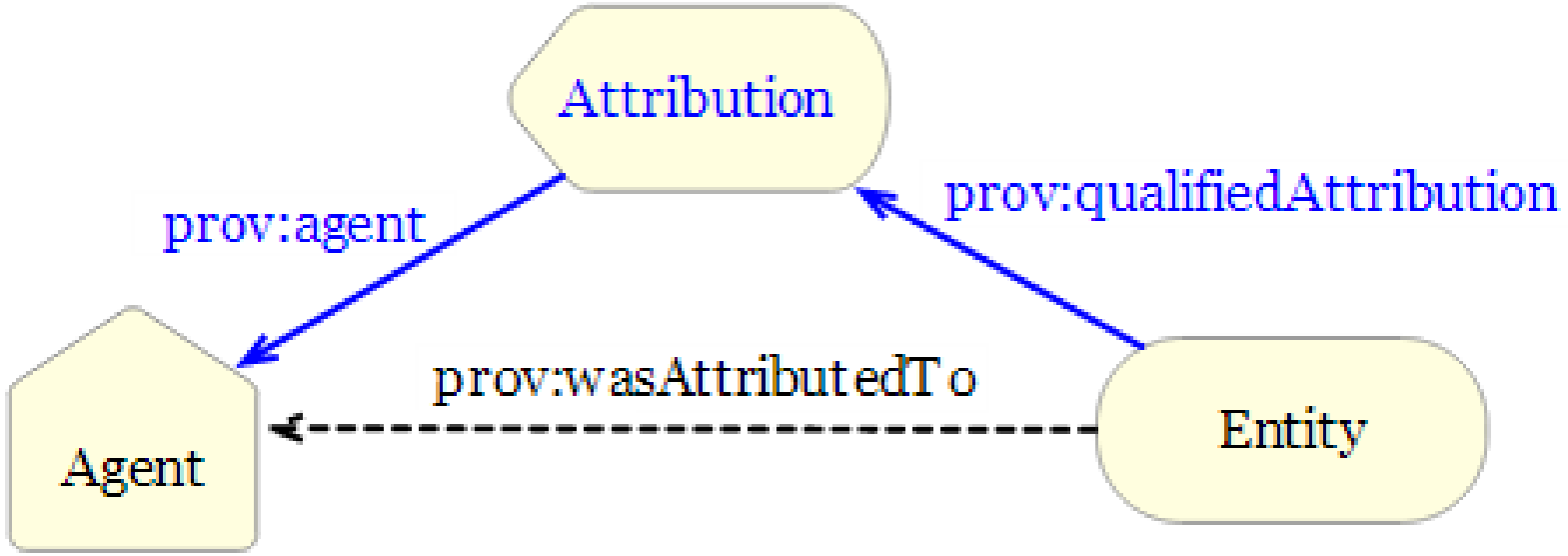
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *actedOnBehalfOf*



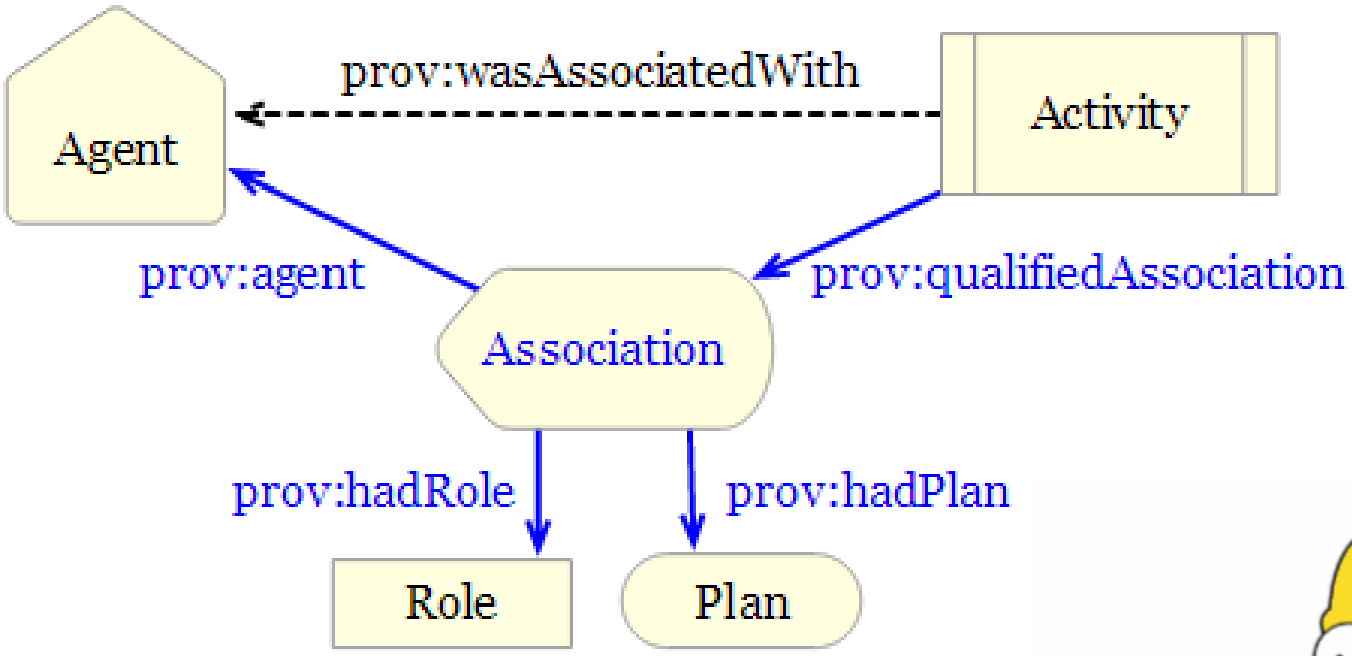
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *wasAttributedTo*



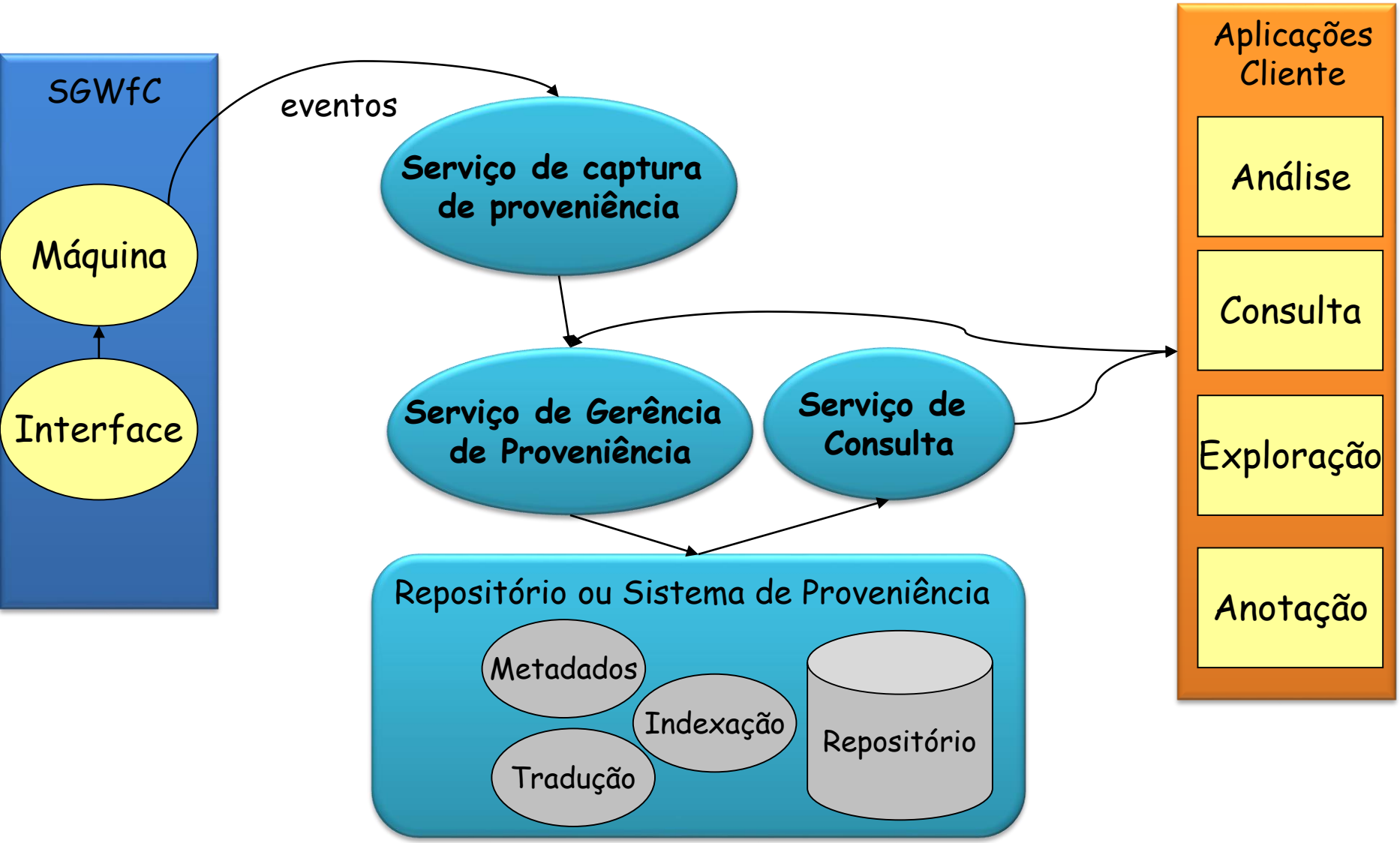
Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Relacionamento *wasAssociatedWith*



Fonte: Sítio W3C PROV www.w3.org/TR/**prov-dm/**

Arquitetura genérica para captura de proveniência



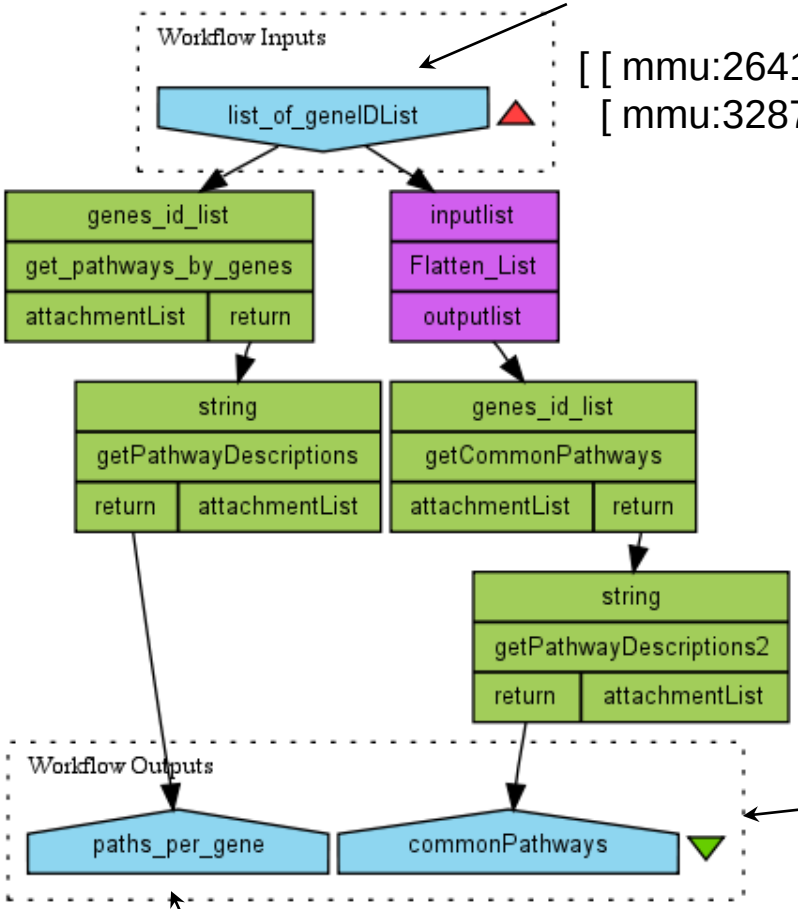
Mecanismos de Captura

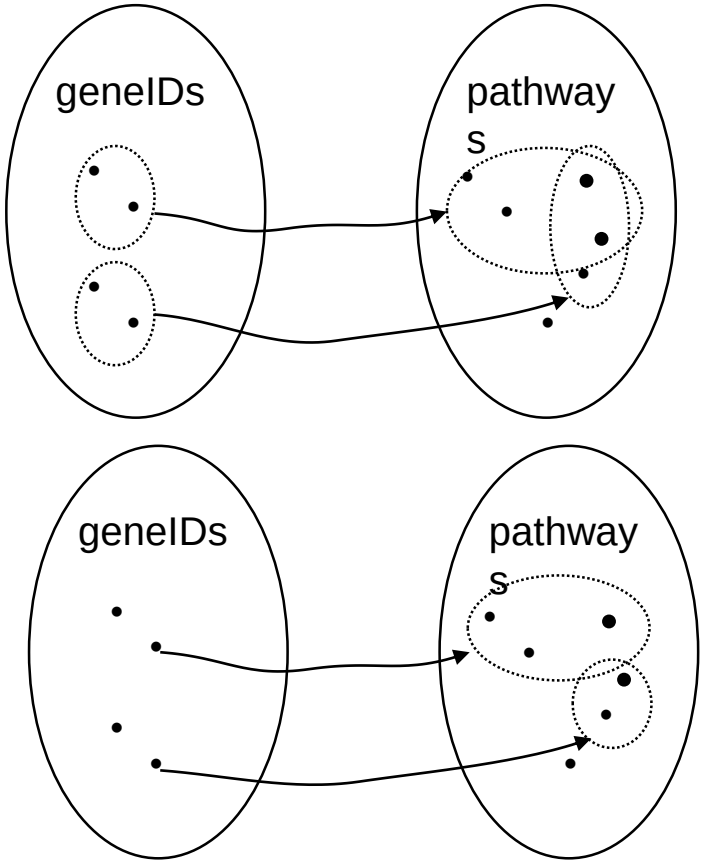
- Atrelados a máquina de execução de *workflows*
  - Taverna
  - Kepler
  - VisTrails
  - SciCumulus
- Desacoplados da máquina de *workflows*
  - PASS
  - PASOA
  - Karma

Proveniência no Taverna

List-structured
KEGG gene ids:

[[mmu:26416],
[mmu:328788]]

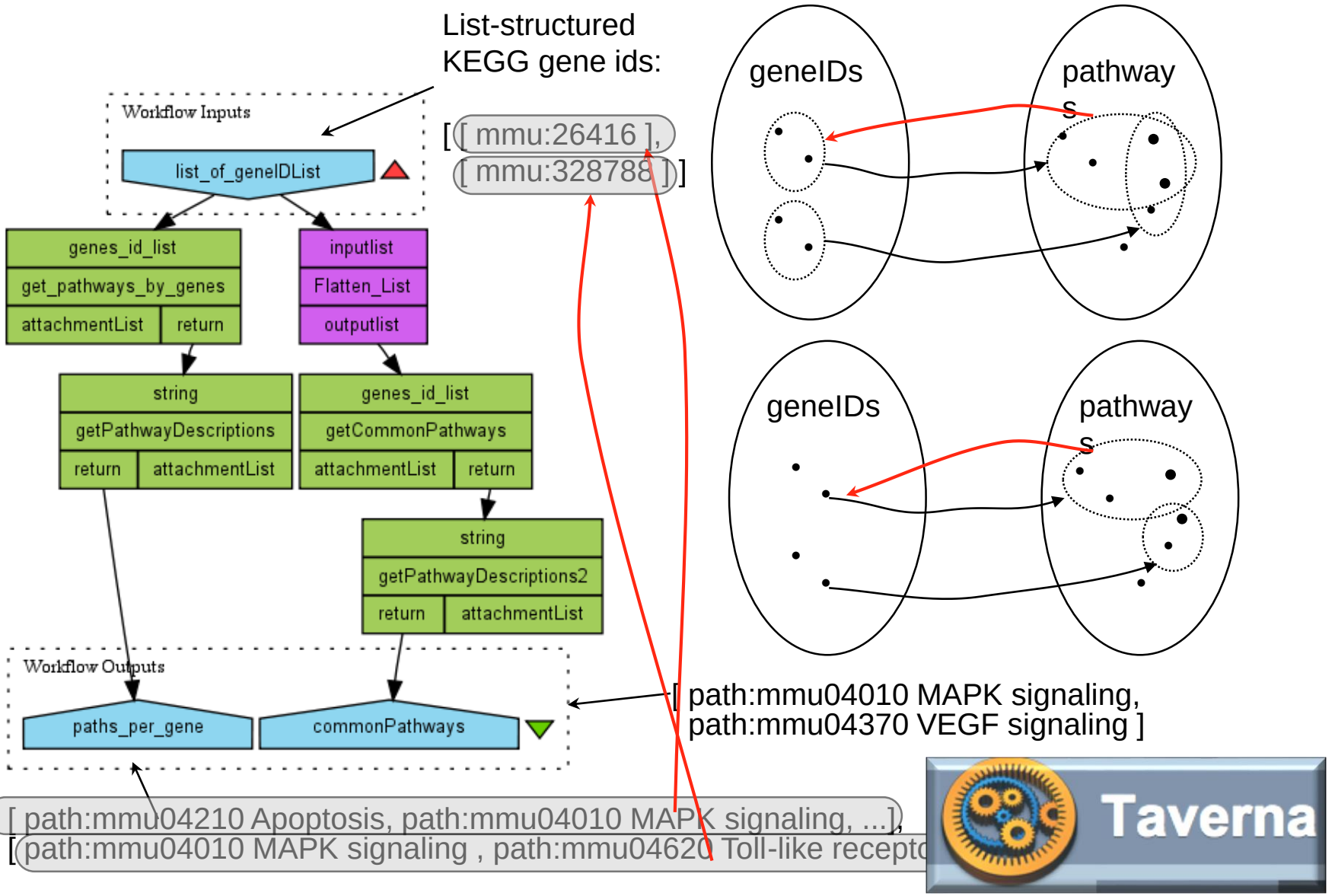




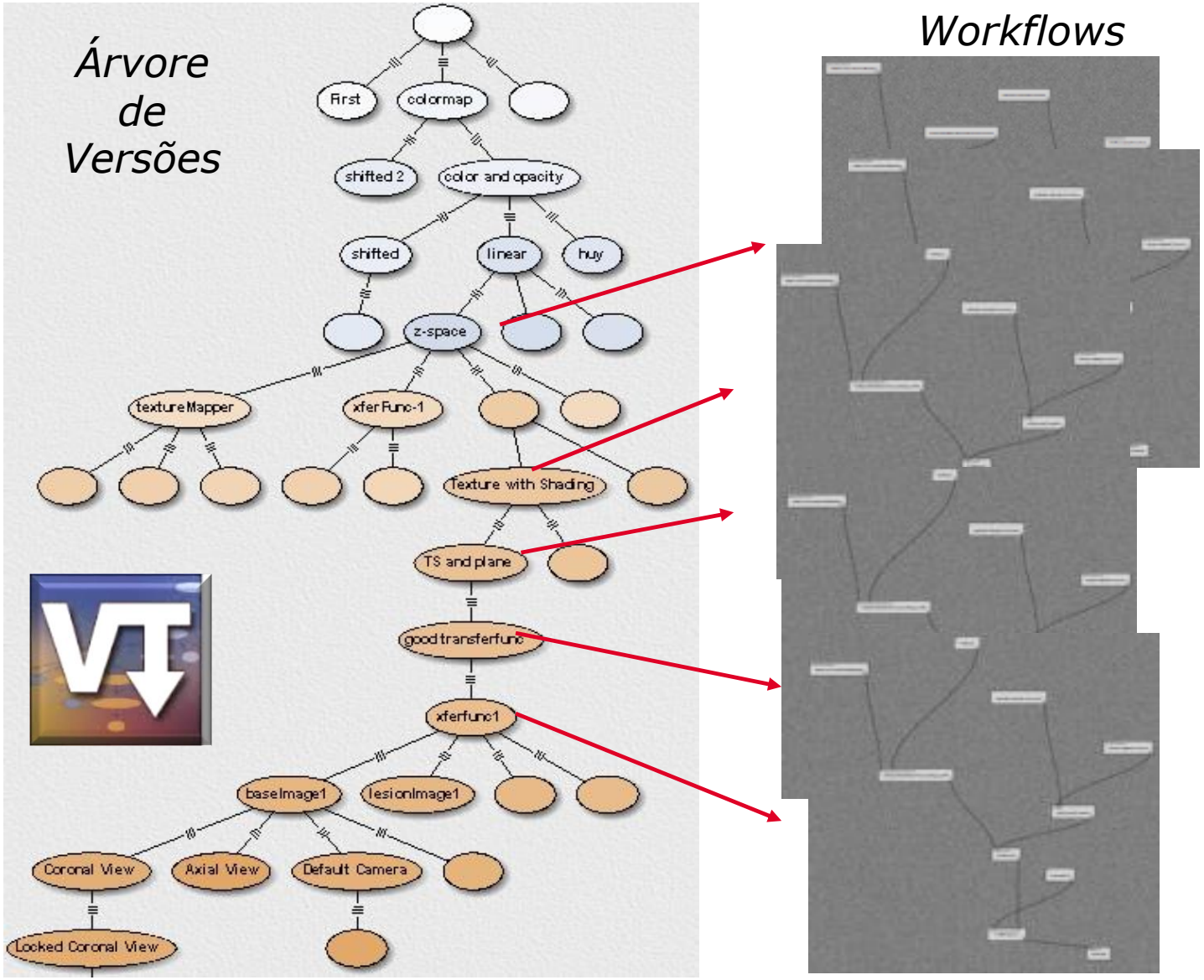
[path:mmu04010 MAPK signaling,
path:mmu04370 VEGF signaling]

[[path:mmu04210 Apoptosis, path:mmu04010 MAPK signaling, ...],
[path:mmu04010 MAPK signaling , path:mmu04620 receptor, ...]]

Proveniência no Taverna

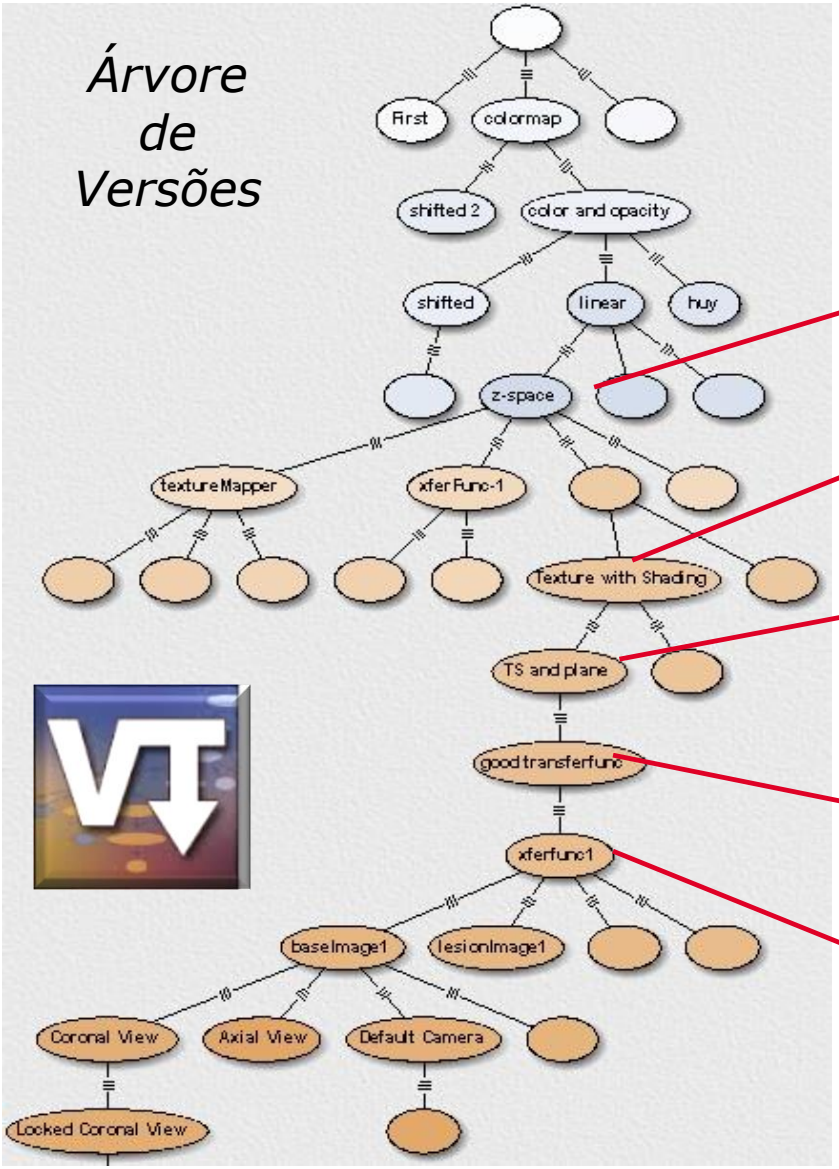


Proveniência no VisTrails

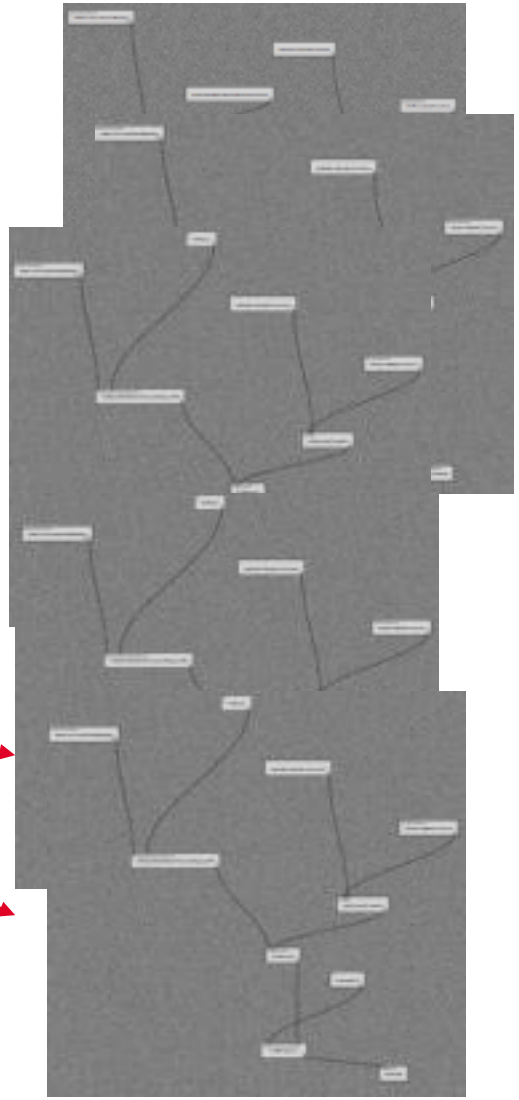


Proveniência no VisTrails

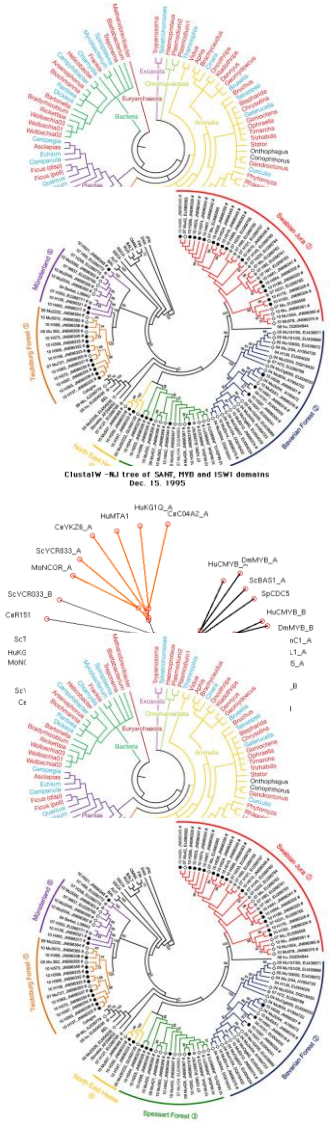
Árvore de Versões



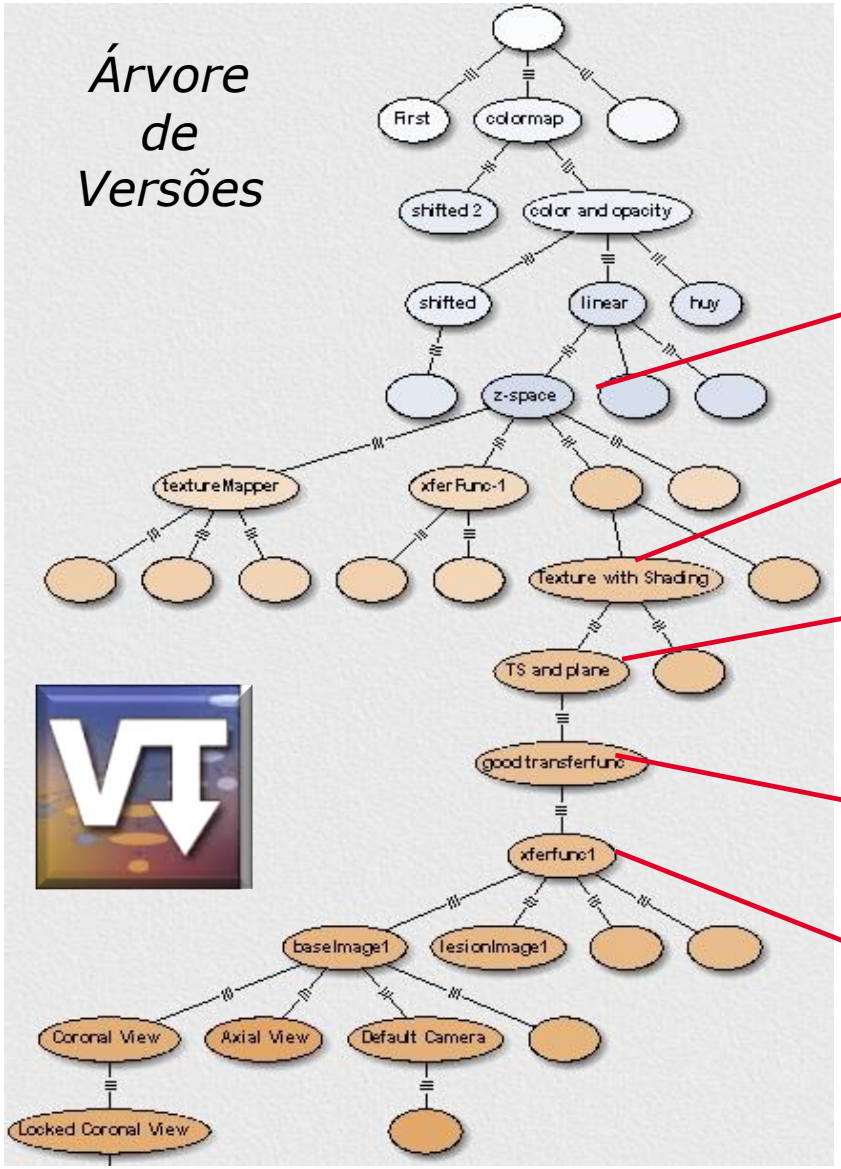
Workflows

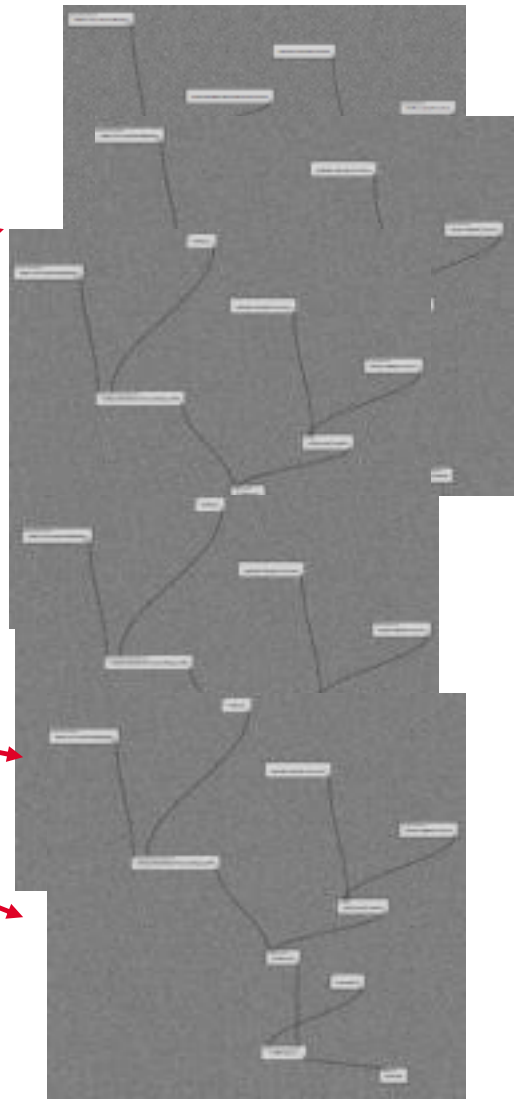


Dados



Proveniência no VisTrails





Usuário

marta

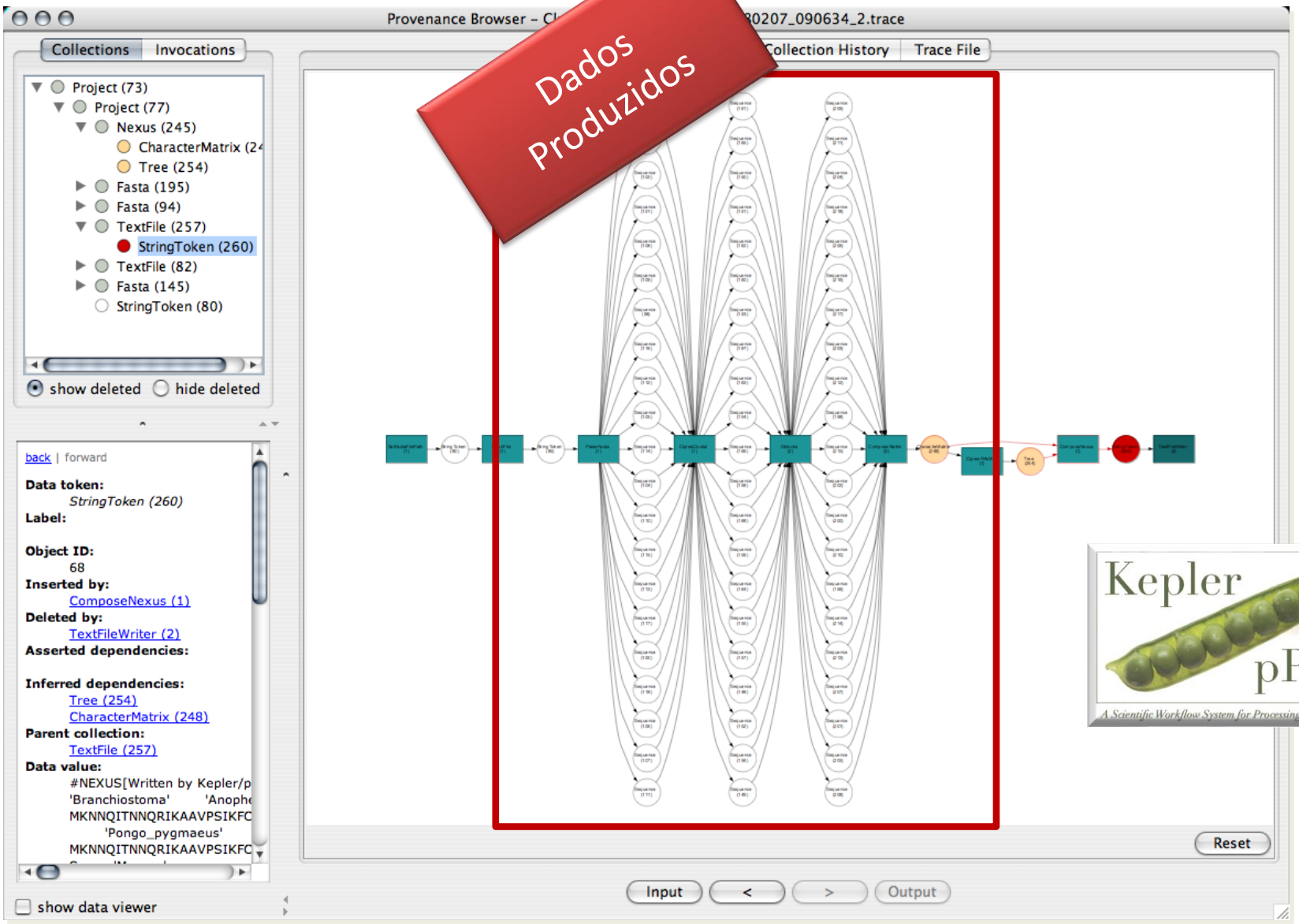
danielcmo

danielcmo

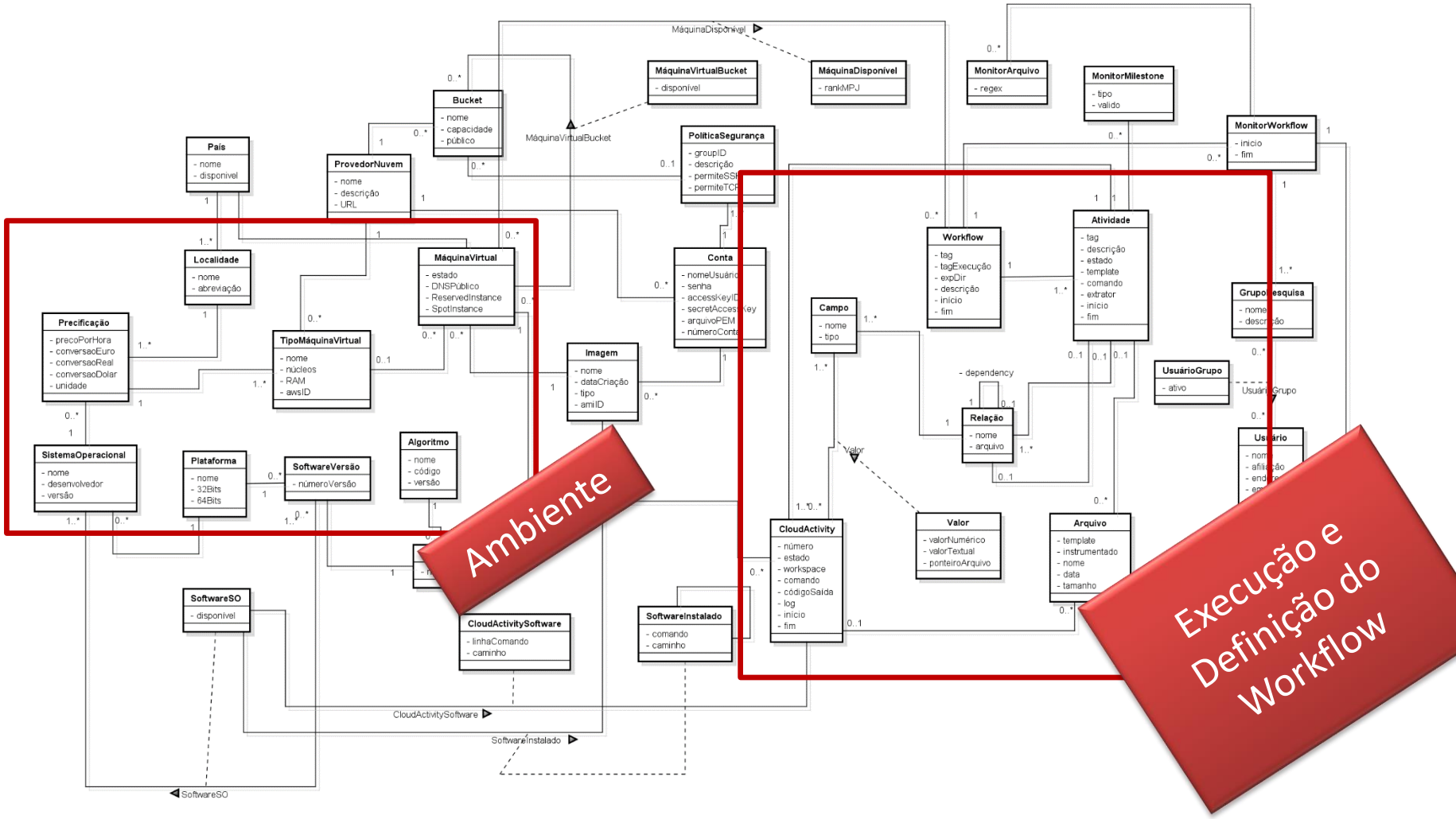
kary

kary

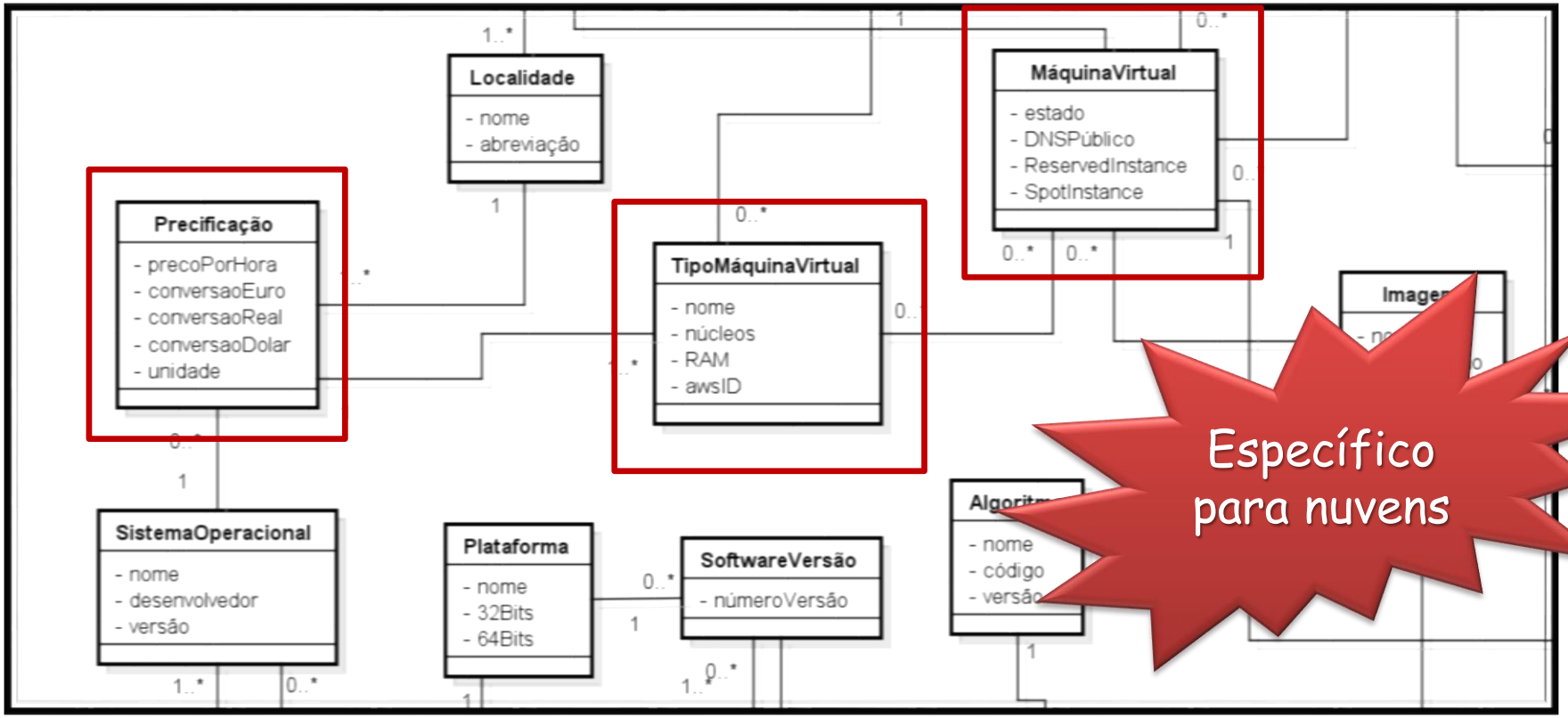
Proveniência no Kepler



Proveniência no Chiron/SciCumulus

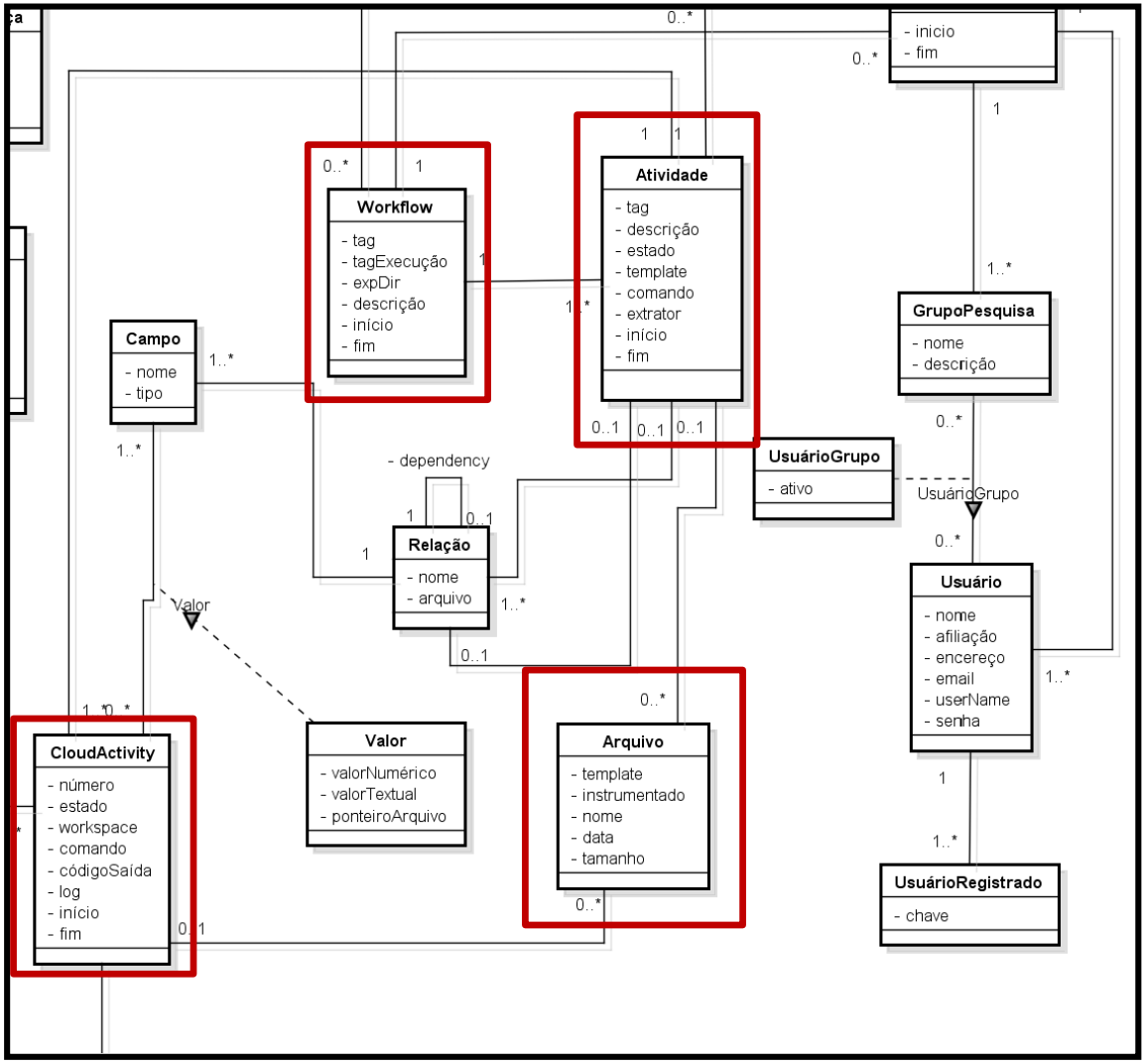


Proveniência no Chiron/SciCumulus



Este tipo de informação pode ser usado para fins além da reprodução e validação

Proveniência no Chiron/SciCumulus



Consultas de Proveniência no SciCumulus

- "Recuperar, por ordem crescente de identificador dos *workflows* as atividades relacionadas a cada *workflow* e as *cloud activities* associadas a cada atividade"

```
SELECT w.tag,
       w.description
       a.tag,
       t.workspace
from hworkflow w, hactivity a, hactivation t
where w.wkfid = a.wkfid
and a.actid = t.actid
order by w.wkfid
```

SciPhy-60003	Phylogeny using RAXML-kalign	raxml	/root/exp/raxml/4/
SciPhy-60003	Phylogeny using RAXML-kalign	raxml	/root/exp/raxml/3/
SciPhy-60003	Phylogeny using RAXML-kalign	raxml	/root/exp/raxml/2/
SciPhy-60003	Phylogeny using RAXML-kalign	raxml	/root/exp/raxml/1/
SciPhy-60003	Phylogeny using RAXML-kalign	readseq	/root/exp/readseq/5/
SciPhy-60003	Phylogeny using RAXML-kalign	readseq	/root/exp/readseq/4/
SciPhy-60003	Phylogeny using RAXML-kalign	readseq	/root/exp/readseq/3/
SciPhy-60003	Phylogeny using RAXML-kalign	readseq	/root/exp/readseq/2/
SciPhy-60003	Phylogeny using RAXML-kalign	readseq	/root/exp/readseq/1/
SciPhy-60003	Phylogeny using RAXML-kalign	mafft	/root/exp/mafft/2/
SciPhy-60003	Phylogeny using RAXML-kalign	mafft	/root/exp/mafft/3/
SciPhy-60003	Phylogeny using RAXML-kalign	mafft	/root/exp/mafft/4/
SciPhy-60003	Phylogeny using RAXML-kalign	mafft	/root/exp/mafft/1/
SciPhy-60003	Phylogeny using RAXML-kalign	mafft	/root/exp/mafft/5/

Consultas de Proveniência no SciCumulus

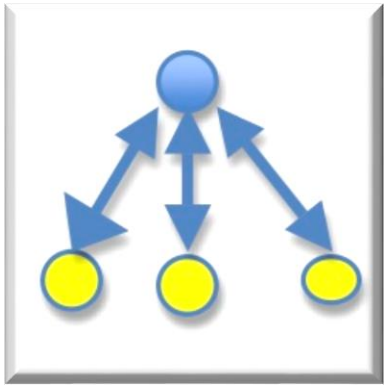
- "Recuperar, por ordem crescente de execuções dos *workflows*, as datas de início e término, *tags* dos *workflows*, bem como o nome de todas as atividades associadas e que não contenham nenhuma *cloud activity* que executou com erro".

```
SELECT w.tag,
        a.tag,
        t.exitstatus,
        t.processor,
        t.workspace,
        t.status,
        t.endtime,
        t.starttime,
        extract ('epoch' from (t.endtime-t.starttime))||',' as duration
from h*workflow* w, h*activity* a, hactivation t
where w.wkfid = a.wkfid
and a.actid = t.actid
and not exists (select * from hactivation a2
                where a2.actid = a.actid
                and a2.exitstatus <> 0)
order by w.wkfid
```

MafftAdaptive	mafft5	0	1	/root/exp/mafft5/47/	FINISHED	2
MafftAdaptive	mafft5	0	1	/root/exp/mafft5/62/	FINISHED	2
MafftAdaptive	mafft5	0	2	/root/exp/mafft5/77/	FINISHED	2
MafftAdaptive	mafft5	0	1	/root/exp/mafft5/105/	FINISHED	2
MafftAdaptive	mafft5	0	2	/root/exp/mafft5/120/	FINISHED	2
MafftAdaptive	mafft5	0	1	/root/exp/mafft5/136/	FINISHED	2
MafftAdaptive	mafft5	0	2	/root/exp/mafft5/149/	FINISHED	2
MafftAdaptive	mafft5	0	2	/root/exp/mafft5/164/	FINISHED	2
MafftAdaptive	mafft5	0	1	/root/exp/mafft5/186/	FINISHED	2
MafftAdaptive	mafft5	0	2	/root/exp/mafft5/200/	FINISHED	2
MafftAdaptive	mafft5	0	1	/root/exp/mafft5/1/	FINISHED	2
MafftAdaptive	mafft5	0	1	/root/exp/mafft5/3/	FINISHED	2
MafftAdaptive	mafft5	0	2	/root/exp/mafft5/2/	FINISHED	2
MafftAdaptive	mafft5	0	1	/root/exp/mafft5/4/	FINISHED	2

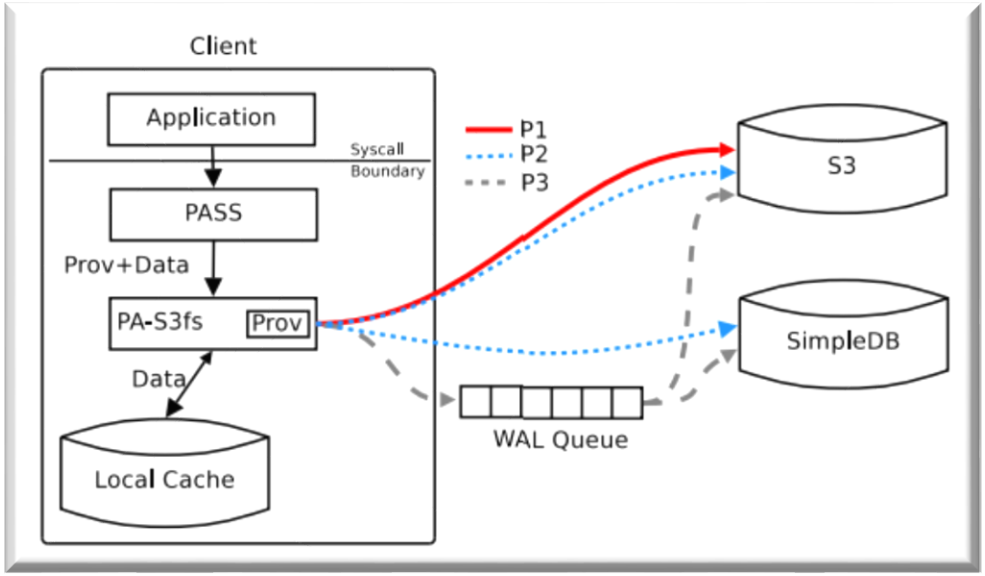
Karma

- http://www.extreme.indiana.edu/karma
- Mecanismo de Captura: cada serviço envia dados de proveniência a um serviço central
- Modelo de Dados:
  - Retrospectiva → grafo representado em XML
  - Prospectiva → BPEL
- Minimiza sobrecarga de captura
  - Desacoplado da máquina de *workflow*



PASS

- http://www.eecs.harvard.edu/syrah/pass/
- Mecanismo de Captura: cada serviço envia dados de proveniência a um serviço central



Proveniência em tempo real

Por que dados de proveniência em tempo real são importantes?

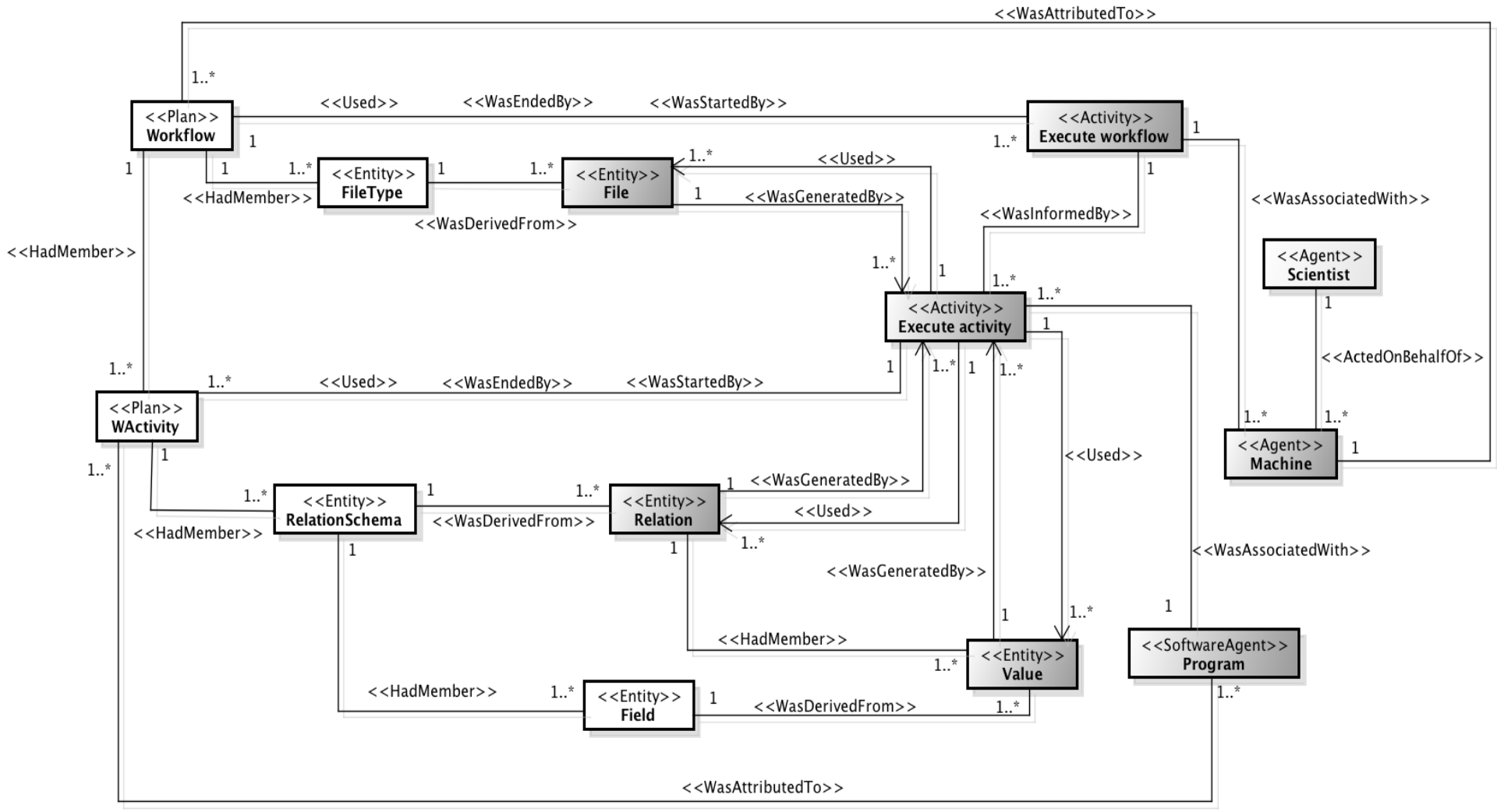
Proveniência

Especialmente em ambientes de nuvem

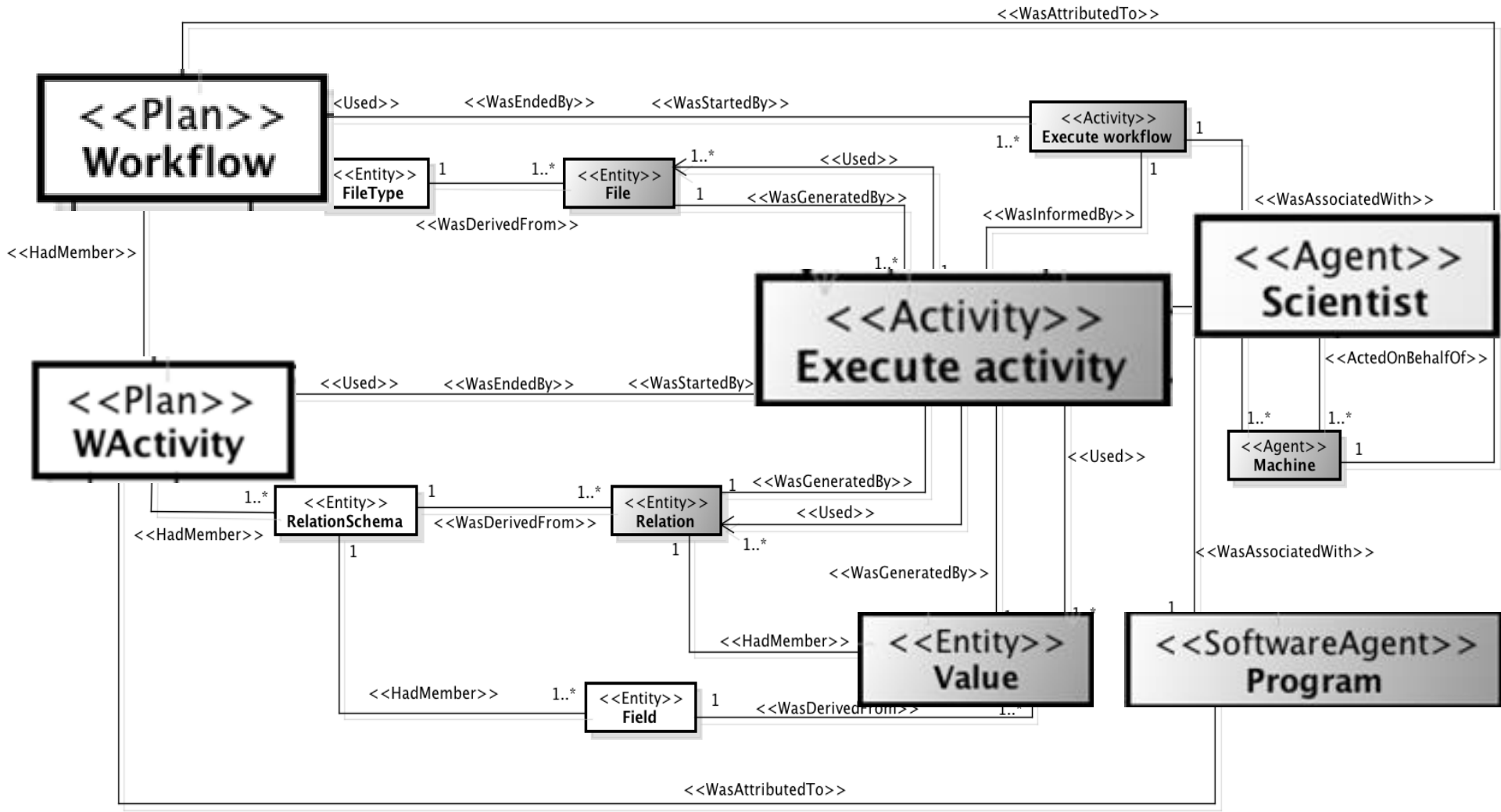
PROV-Wf

- Extensão do PROV para representação de proveniência em *workflows* científicos
- Relaciona elementos do modelo estendido com os elementos do PROV por meio de estereótipos da UML
- Proposto no *workshop* BigProv do EDBT/ICDT 2013

PROV-Wf


<<WasAttributedTo>>
<<Plan>>
Workflow
<<Used>>
<<WasEndedBy>>
<<WasStartedBy>>
<<Activity>>
Execute workflow
<<Entity>>
FileType
<<HadMember>>
<<Entity>>
File
<<WasDerivedFrom>>
<<WasGeneratedBy>>
<<WasInformedBy>>
<<WasAssociatedWith>>
<<Agent>>
Scientist
<<ActedOnBehalfOf>>
<<Activity>>
Execute activity
<<Plan>>
WActivity
<<Agent>>
Machine
<<Entity>>
RelationSchema
<<Entity>>
Relation
<<Entity>>
Value
<<SoftwareAgent>>
Program
<<Entity>>
Field

PROV-Wf

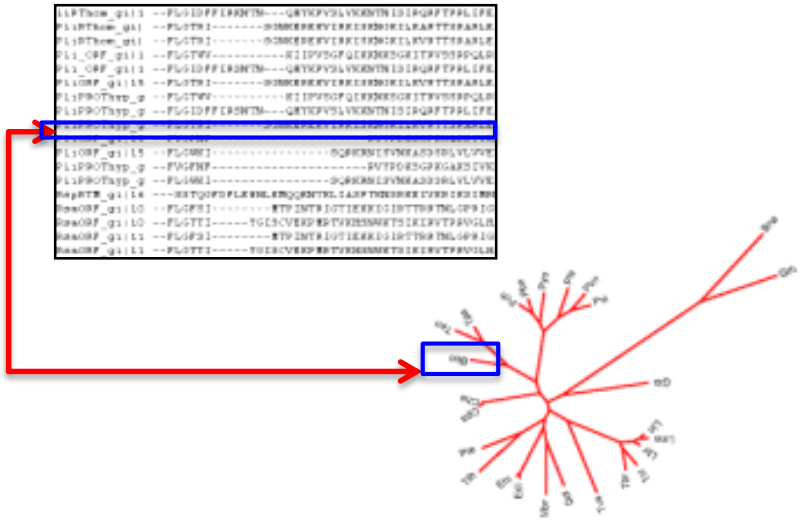

<<WasAttributedTo>>
<<Plan>>
Workflow
<<Used>>
<<WasEndedBy>>
<<WasStartedBy>>
<<Activity>>
Execute workflow
<<Entity>>
FileType
<<Entity>>
File
<<WasDerivedFrom>>
<<WasGeneratedBy>>
<<WasInformedBy>>
<<WasAssociatedWith>>
<<HadMember>>
<<Agent>>
Scientist
<<Activity>>
Execute activity
<<Plan>>
WActivity
<<ActedOnBehalfOf>>
<<Agent>>
Machine
<<Entity>>
RelationSchema
<<Entity>>
Relation
<<Entity>>
Value
<<SoftwareAgent>>
Program
<<Entity>>
Field
1
1..*

Mapeamento no PROV-Wf

VisTrails	*SciCumulus*	*PROV-Wf*
VisTrail.ID	Workflow.ExecTag	Execute Workflow.ID
VisTrail.Name	Workflow.Tag	Workflow. Name
Module.Name	Activity.Tag	WActivity. Name
Module.ID	Activation.ID	Execute Activity.ID
Parameter.Alias	Field.Name	Field.Name
Parameter.Type	Field.Type	Field.Type
Parameter.Val	Value.Val	Value.Val

Uso do PROV-Wf

SELECT f1.name, v1.value, f2.name, v2.value
FROM workflow w, activity a, executeActivity ea, relationschema rs, relation r,
field f1, field f2, Value v1, Value v2
WHERE a.wkf_id = w.id **AND** ea.act_id = a.id
AND r.rel_id = rs.id **AND** r.ea_id = ea.id
AND f1.rel_id = rs.id **AND** f2.rel_id = rs.id
AND v1.uf_id = f1.id **AND** v2.uf_id = f2.id
AND rs.name = “Relation MSA”
AND f1.name = “BIO_SEQ”
AND f2.name = “TAXON”
AND a.name = %PARAM_ACT_NAME%
AND w.tag = %PARAM_WF_NAME%



Roteiro do Tutorial

- Motivação
- Workflows Científicos
- Nuvens de Computadores
- Proveniência de Dados
- **Máquinas de Workflow para Nuvem**
- Aplicação de Proveniência em e-Science
- Demo

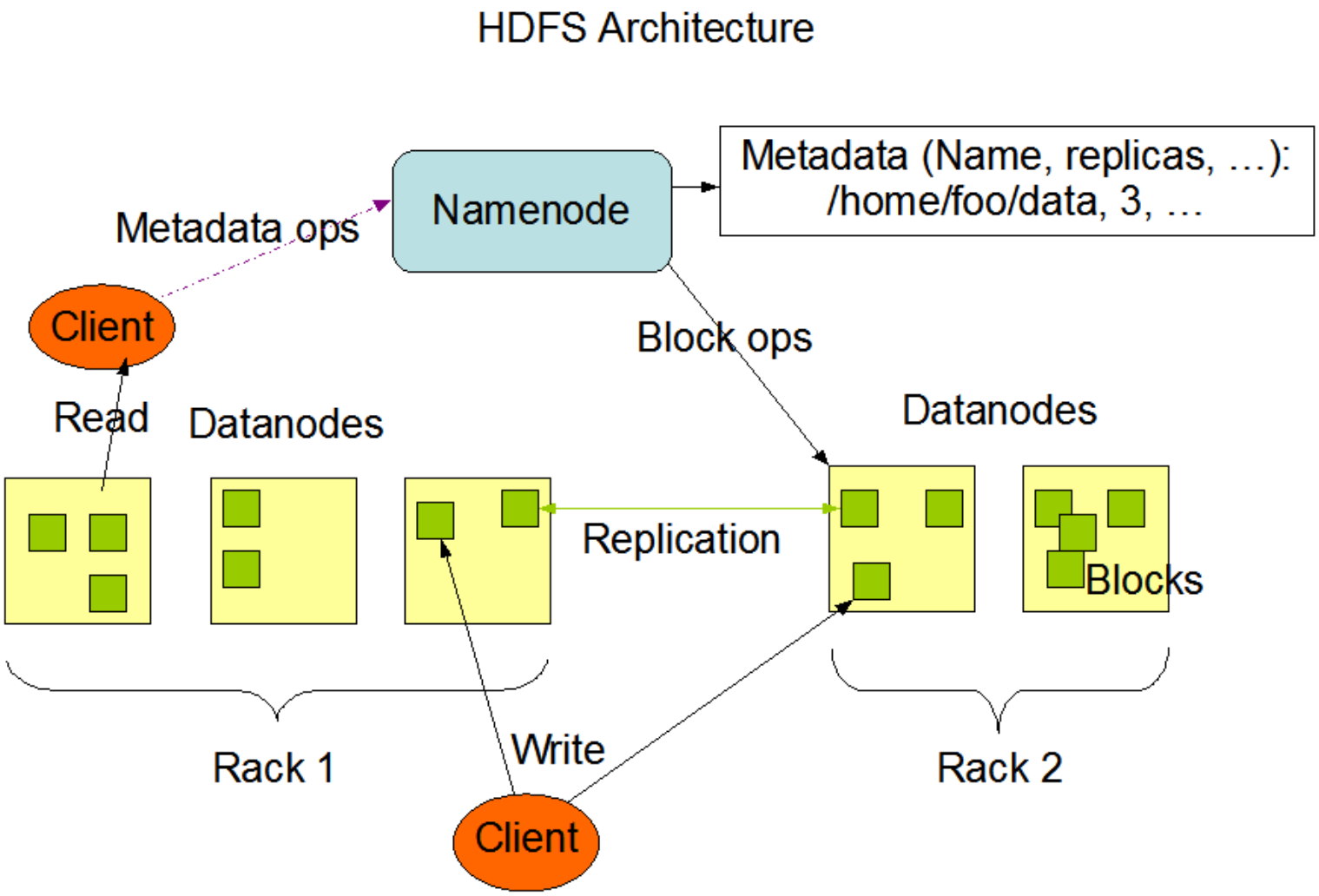
Máquinas baseadas em Hadoop

- Framework para execução de aplicações em grandes clusters (virtuais ou não)
- Fornece funcionalidades tanto de confiabilidade e movimentação de dados.
- Implementa o paradigma computacional chamado map / reduce
  - Cada aplicação é "dividida" em muitos pequenos fragmentos, cada um dos quais podem ser executadas ou reexecutado em qualquer nó no cluster.

Hadoop

- Desacoplado do conceito de workflow
- Escalonamento não leva em consideração a estrutura do workflow
- Não captura proveniência de forma nativa
  - Complementos para captura são necessários

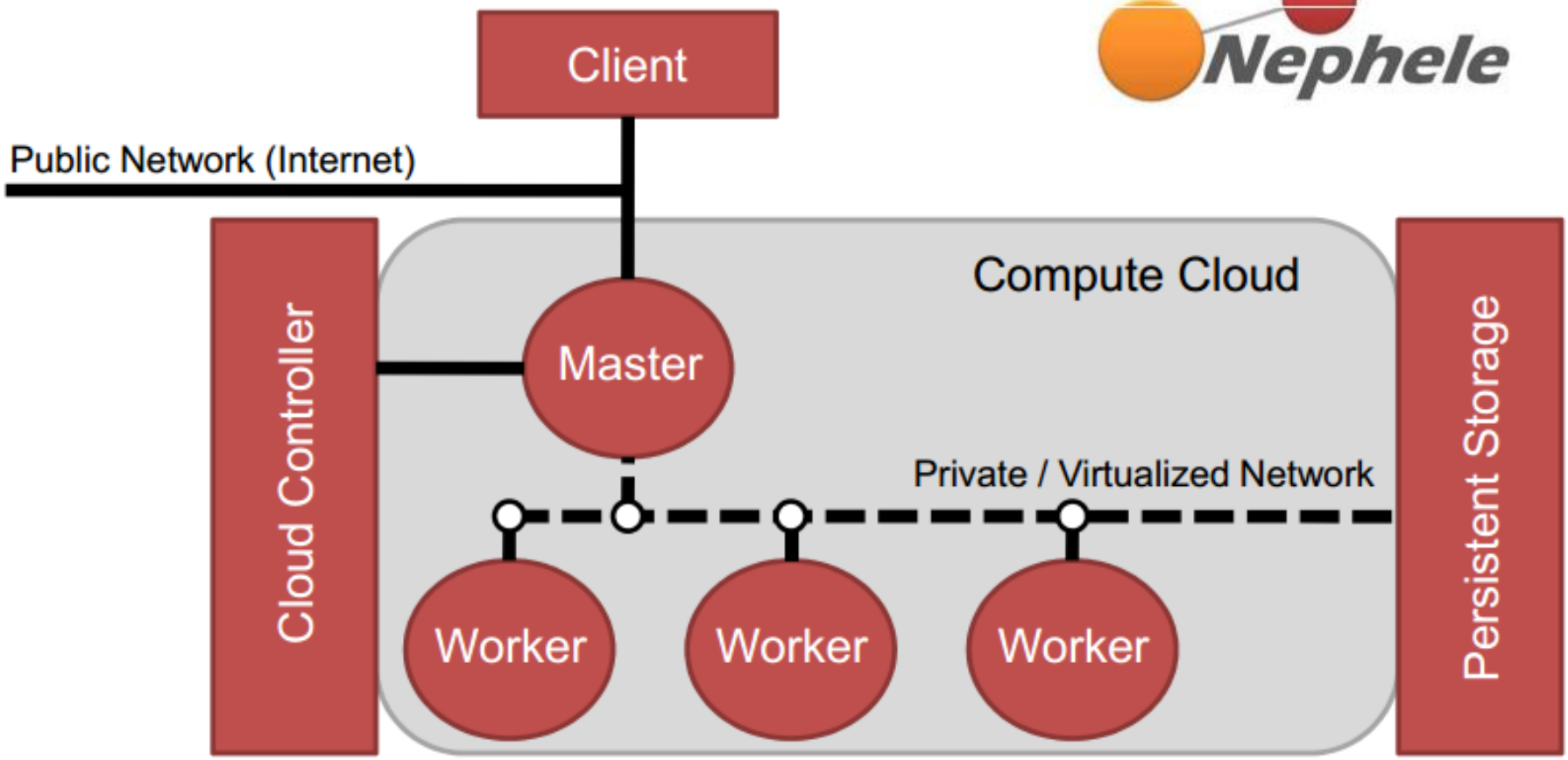
Arquitetura do Hadoop



Nephele

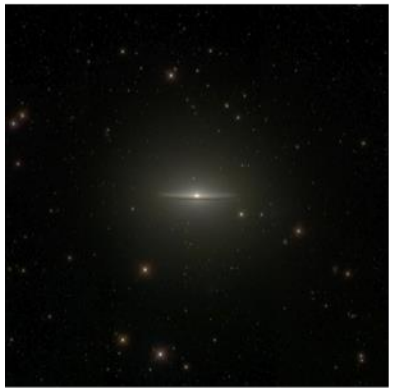
- Arcabouço para processamento paralelo em nuvens
- Aloca e desaloca máquinas virtuais dependendo do comportamento do Job
- Não possui proveniência associada
- Paralelização explícita
  - Tarefas são designadas como "parallelizable"

Nephele


Client
Public Network (Internet)
Cloud Controller
Compute Cloud
Master
Private / Virtualized Network
Worker
Worker
Worker
Persistent Storage

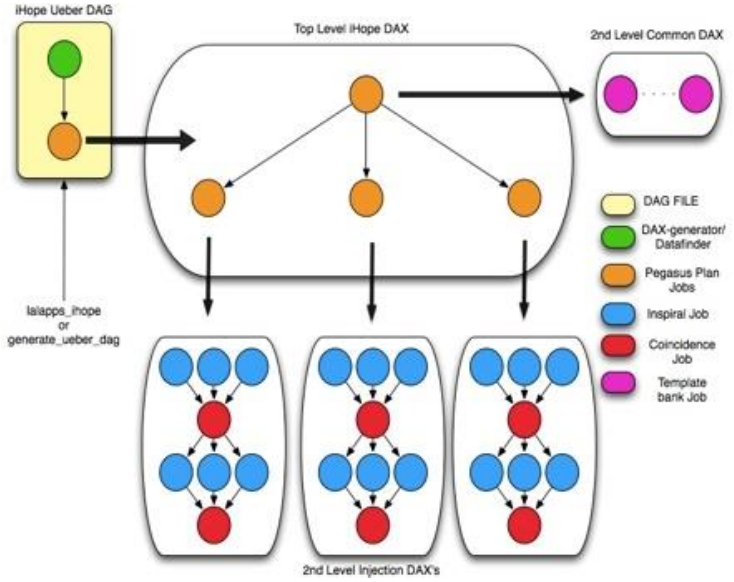
Pegasus

- Financiado pela NSF/OCI em colaboração com o grupo do escalonador CONDOR de UW Madison
- Distribui atividades em paralelo em ambientes distribuídos
- Captura os dados de proveniência
- Escalabilidade
  - Big Data (kB to TB)
  - Quantidade de tarefas (1…10^6 tarefas)

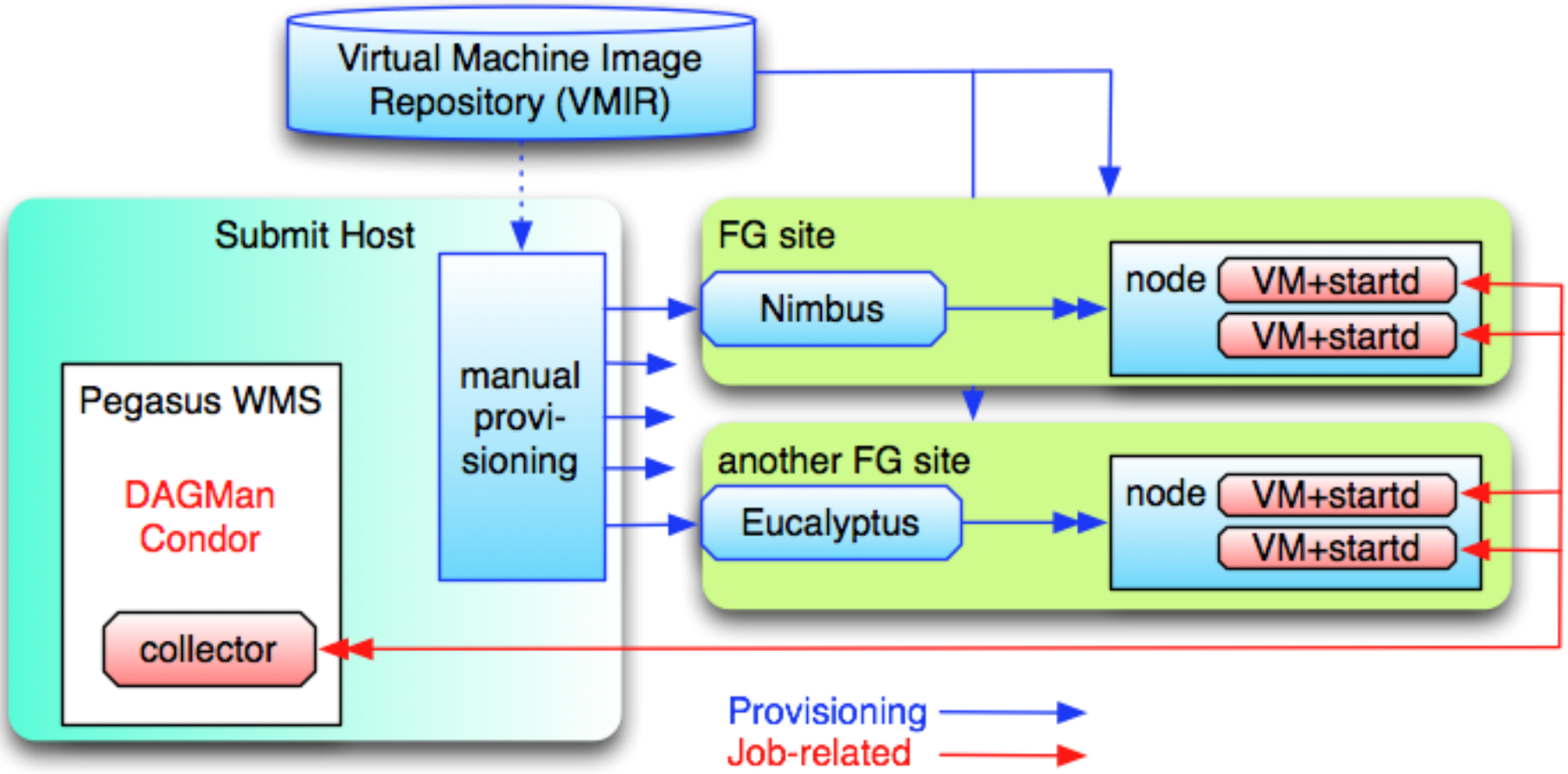


Pegasus

- Re-executa tarefas que falharam automaticamente
- Pode executar nos mais diversos ambientes
  - Laptop
  - campus cluster
  - grid
  - **nuvem**



Pegasus



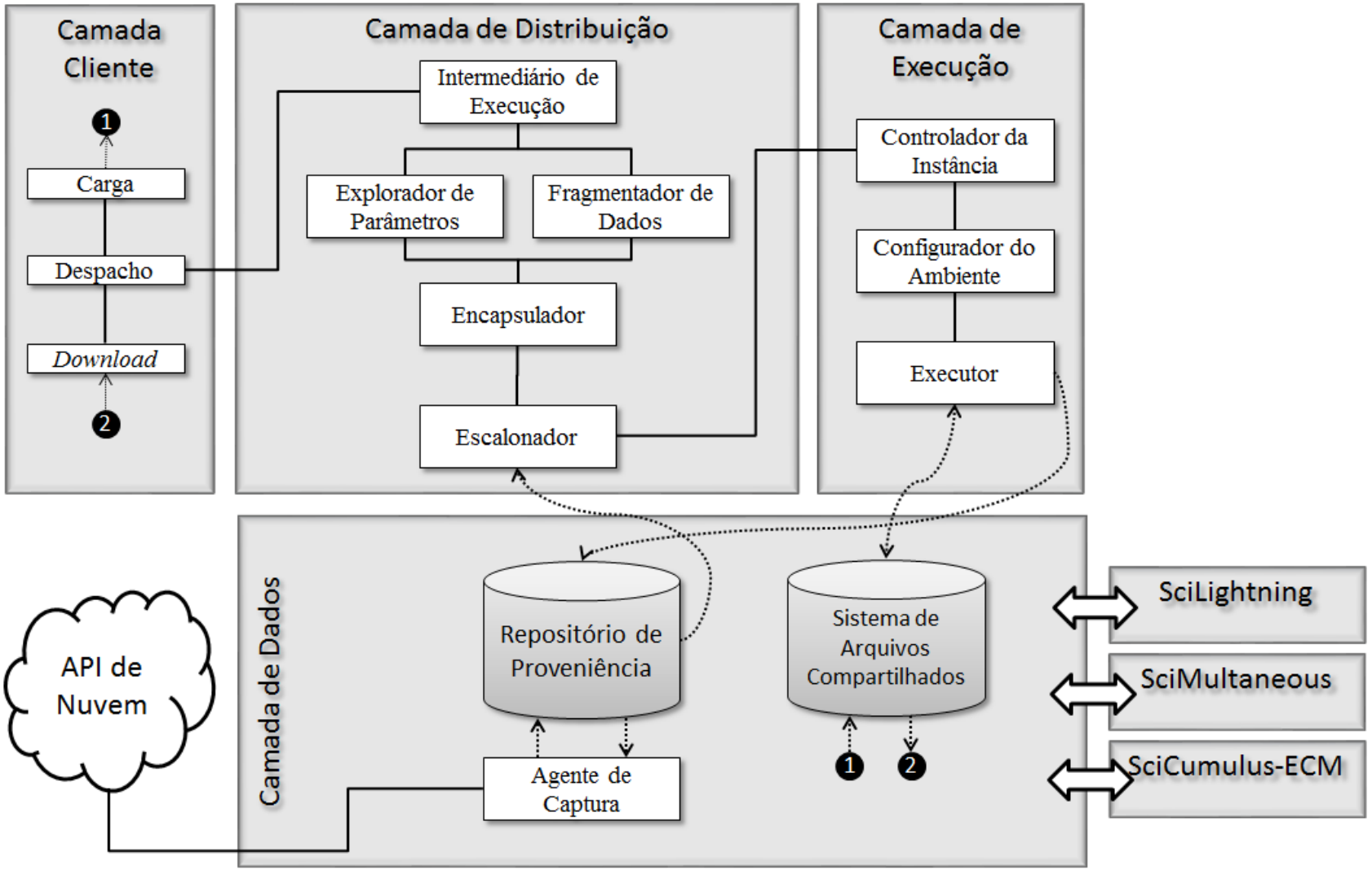
Fonte: pegasus.isi.edu/presentations/

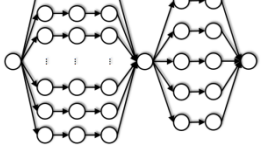
SciCumulus

- Uma máquina para execução de *workflows* para nuvem que tem como objetivo conectar o SGWfC com o ambiente de nuvem para oferecer apoio à execução de *workflows* científicos
  - *Middleware* Hydra
  - Motor de execução Chiron

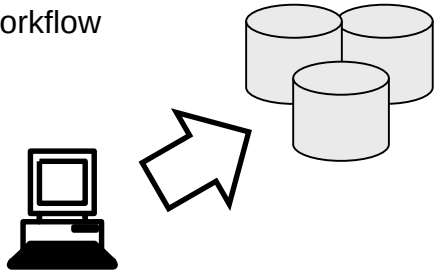
OLIVEIRA, D.; OGASAWARA, E.; BAIAO, F.; MATTOSO, M. L. Q. . **SciCumulus: A Lightweight Cloud Middleware to Explore Many Task Computing Paradigm in Scientific Workflows**. In: The 3rd IEEE CLOUD 2010, Miami. Proc. of the 3rd IEEE CLOUD, 2010. p. 378-385

Arquitetura





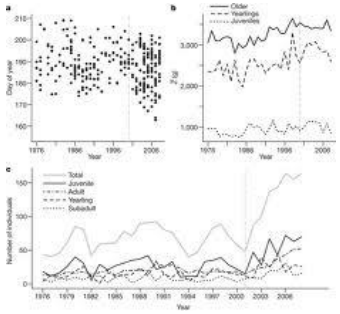
especificação do workflow



transferência de dados

para a nuvem

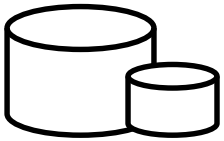
validação, verificação
e publicação dos
resultados



composição

descoberta

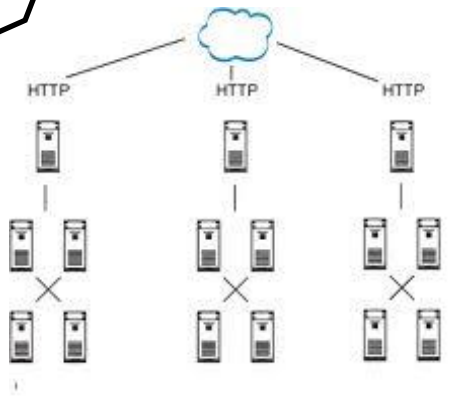
análise



Dados de
Proveniência

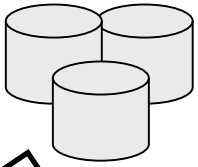
configuração

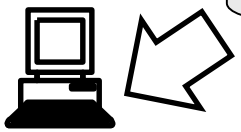
Configuração do cluster
virtual



redimensionamento

execução

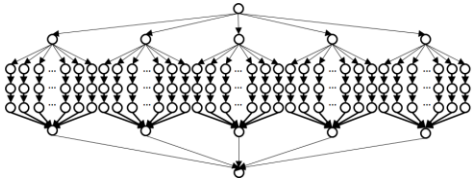




download de dados

da nuvem

monitoramento da execução
na nuvem



execução distribuída e
paralela do workflow

Modelo Algébrico

- O SciCumulus segue o modelo algébrico proposto por Ogasawara *et al.* (2011)
  - Atividades consomem relações
  - Cada relação é composta de tuplas e atributos (parâmetros das atividades)
  - Execuções de atividades consomem tuplas

- Os operadores da álgebra Invocam programas:
  - **Map (1:1)**
  - SplitMap (1:n)
  - Reduce (n:1)
  - Filter (1:0-1)

OGASAWARA, E. ; DIAS, J. ; OLIVEIRA, D. ; PORTO, F. ; VALDURIEZ, P. ; MATTOSO, M. L. Q. . **An Algebraic Approach for Data-Centric Scientific Workflows**. Proceedings of the **VLDB** Endownment, v. 4, p. 1328-13369, 2011.

Relações como o modelo de dados para consumo e produção

- Relações são definidas como um conjunto de tuplas de tipos primitivos (inteiro, string, etc) ou tipos complexos (e.g. ponteiros para arquivos)
- Exemplo: *R(R)*

RID	**CaseStudy**	**sdat**	**ddat**
1	U-125	U-125S.DAT	U-125D.DAT
1	U-127	U-127S.DAT	U-127D.DAT
2	U-129	U-129S.DAT	U-129D.DAT

- *R = (RID: Integer, CaseStudy: String; SDat: FileRef, DDat: FileRef)*

Split Map (SplitMap)
T ← SplitMap(Y, a, R)

R

RID	**RdZip**
1	Project1.zip
2	Project2.zip

T ← SlipMap(extractRD, 'RdZip', R)

T

RID	**Study**	**sdat**	**ddat**
1	U-125	U-125S.DAT	U-125D.DAT
1	U-127	U-127S.DAT	U-127D.DAT
2	U-129	U-129S.DAT	U-129D.DAT

Reduce Activity (Reduce)
$T \leftarrow Reduce(Y, g_A, R)$

R

RID	Study	SsSai	DdSai	MEnv
1	U-125	U-125Ss.SAI	U-125Dd.SAI	U-125.ENV
1	U-127	U-127Ss.SAI	U-127Dd.SAI	U-127.ENV
2	U-129	U-129Ss.SAI	U-129Dd.SAI	U-129.ENV

T ← Reduce(CompressRD, {'RID'}, R)

T

RID	RdResultZip
1	ProjectResult1.zip
2	ProjectResult2.zip

SRQuery Activity
$T \leftarrow SRQuery(qry, R)$

R

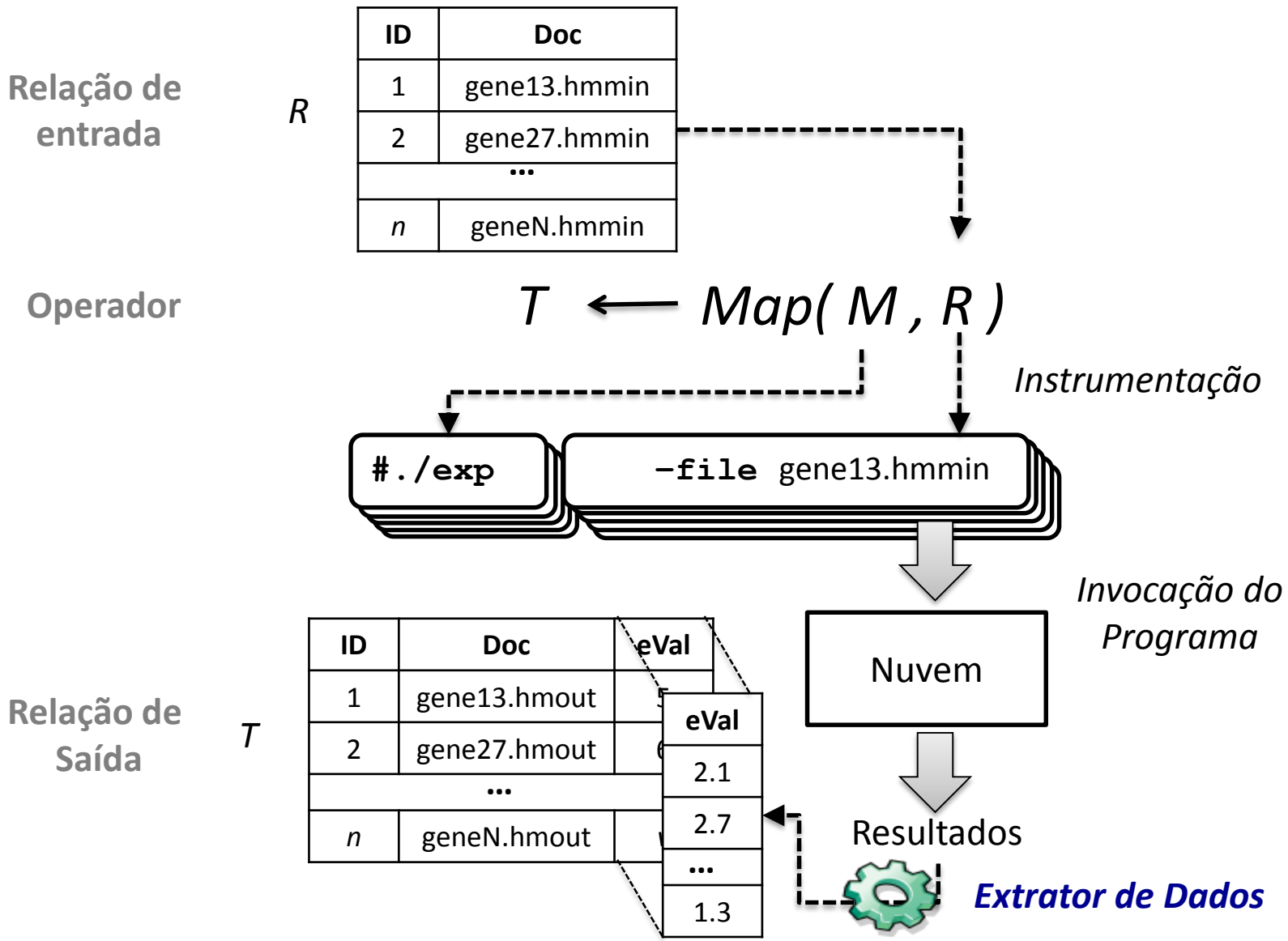
RID	Study	SsSai	Curvature
1	U-125	U-125Ss.SAI	1.5
1	U-126	U-126Ss.SAI	0.9
1	U-127	U-127Ss.SAI	1.2

$$T \leftarrow SRQuery(\pi_{RID,\ Study,\ SsSai,\ Curvature}(\sigma_{Curvature>1}(R)),\ R)$$

T

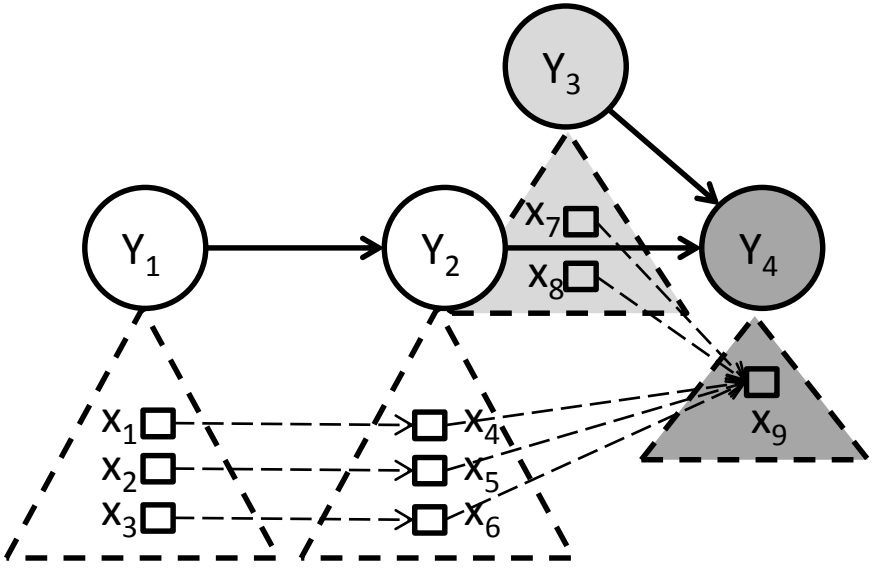
RID	Study	SsSai	Curvature
1	U-125	U-125Ss.SAI	1.5
1	U-127	U-127Ss.SAI	1.2

Execução baseada em relações



Estratégias de Execução

- *First Tuple First (FTF)*
- *First Activity First (FAF)*



FTF:

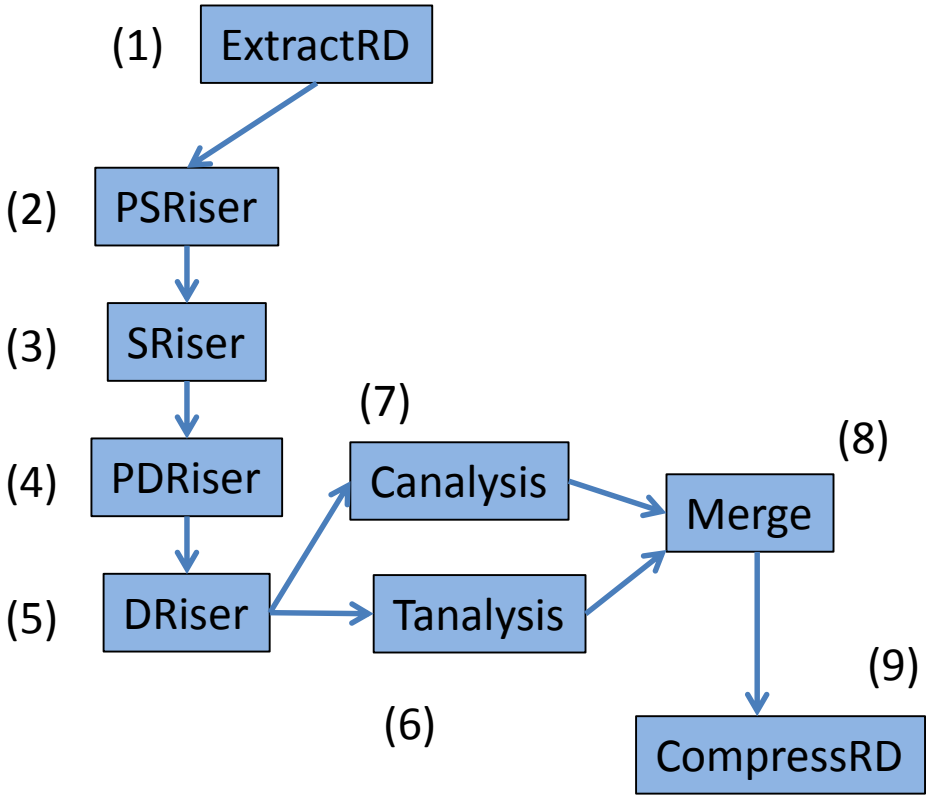
{<x_1, x_4>, <x_2, x_5>, <x_3, x_6>}

FAF:

{<x_1>,<x_2>,<x_3>,
<x_4>,<x_5>,<x_6>}

Workflow como uma expressão da álgebra

Workflow



Expressões algébricas

$T_1 \leftarrow SplitMap(ExtractRD, R_1)$

$T_2 \leftarrow Map(PSRiser, T_1)$

$T_3 \leftarrow Map(SRiser, T_2)$

$T_4 \leftarrow Map(PDRiser, T_3)$

$T_5 \leftarrow Map(DRiser, T_4)$

$T_6 \leftarrow Filter(Tanalysis, T_5)$

$T_7 \leftarrow Filter(Canalysis, T_5)$

$T_8 \leftarrow SRQuery(T_6 \bowtie T_7, \{T_6, T_7\})$

$T_9 \leftarrow Reduce(CompressRD, T_8)$

Detalhes de Implementação

- Java Versão 6 *Update* 15
- HSQLDB
- *Drivers* JDBC
- MPJ – arcabouço de paralelismo em Java
- PostgreSQL versão 8.4.6

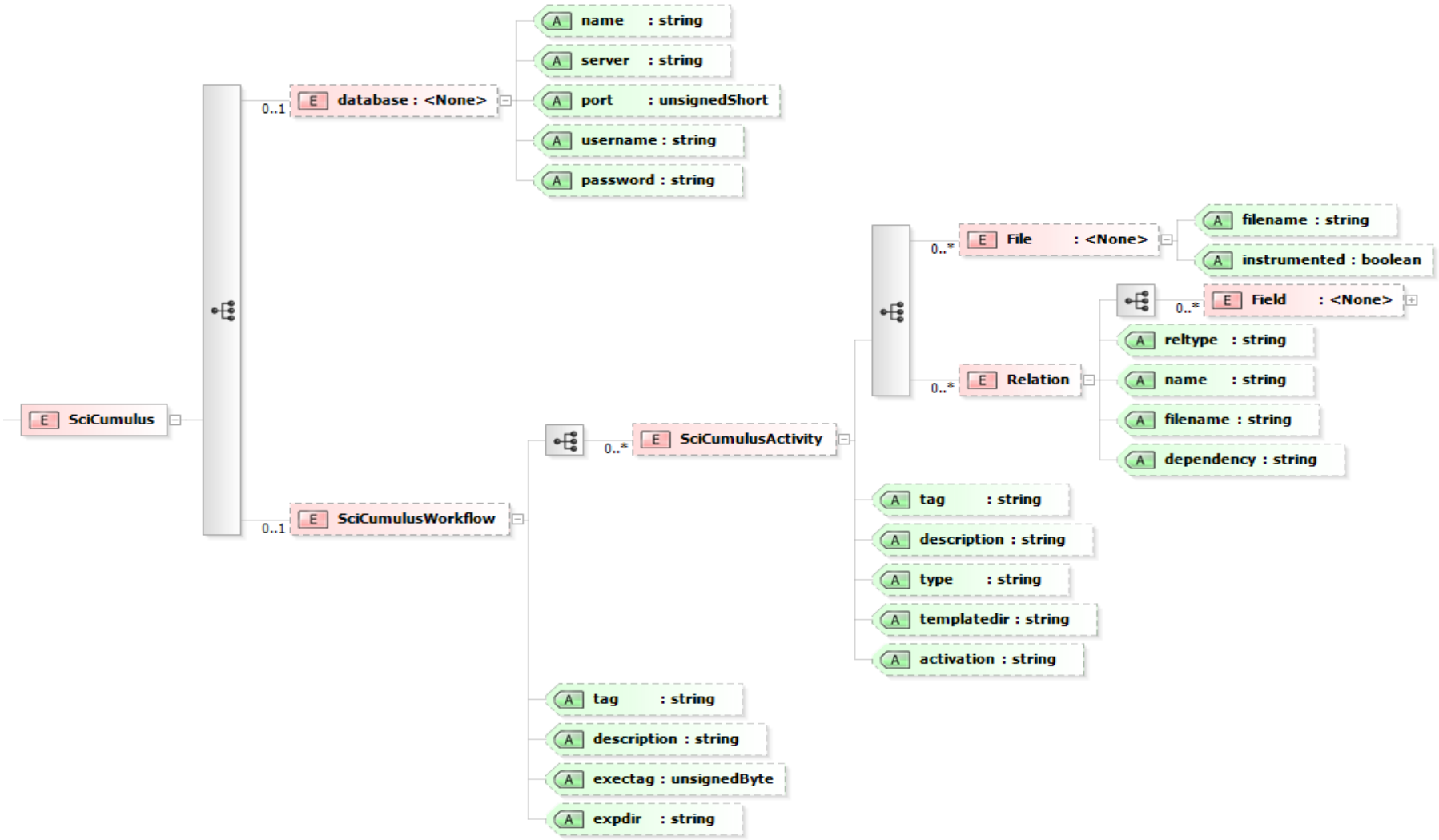
- Código disponibilizado em:
  - https://sourceforge.net/projects/scicumulus/

Representação dos *Workflows*

```
<?xml version="1.0" standalone="no"?>
<SciCumulus>
   <database name="scicumulus_adaptive" server="mp4-4.dyndns.info"
      port="5432" username="scicumulus" password="**********" />
   <SciCumulusWorkflow tag="filogenia" description="This is a test using anflex."
      exectag="Experimento Kary/200 organismos - Adaptive" expdir="/root/exp">
      <SciCumulusActivity tag="mafft" description="mafft" type="MAP"
            templatedir="/root/exp/template_mafft" activation="experiment.cmd">
         <Relation reltype="Input" name="A" filename="parameter.txt" />
         <Field name="NAME" type="string" />
         <Field name="FASTA_FILE" type="string" />
         <Relation reltype="Output" name="C" filename="output_mafft.txt" />
         <Field name="NAME" type="str
         <Field name="NUM_SEQ" type=
         <Field name="FASTA_FILE" ty
         <File filename="experiment.cmd
      </SciCumulusActivity>
   </SciCumulusWorkflow>
</SciCumulus>
```

```
NAME;FASTA_FILE
G1;ORTHOMCL2033
G2;ORTHOMCL1895
G3;ORTHOMCL2034
G4;ORTHOMCL1896
G5;ORTHOMCL2035
G6;ORTHOMCL1897
G7;ORTHOMCL2036
G8;ORTHOMCL1898
G9;ORTHOMCL2037
G10;ORTHOMCL1899
```

Representação dos *Workflows*

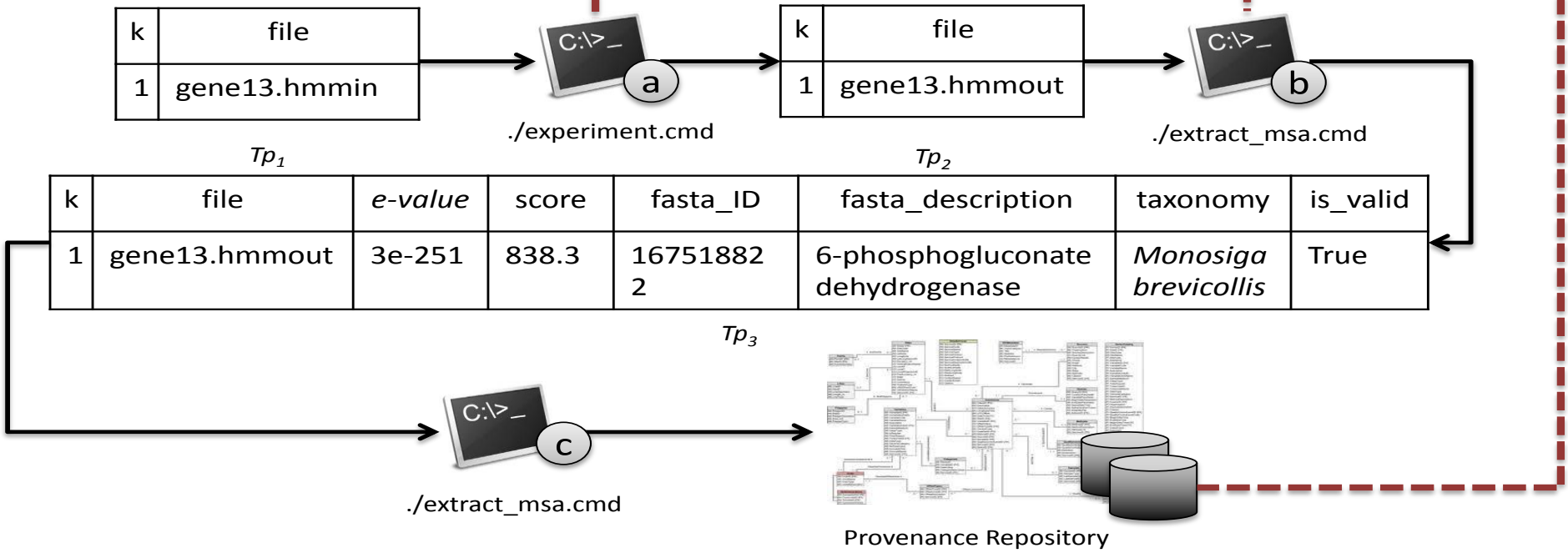


Componente extrator de dados

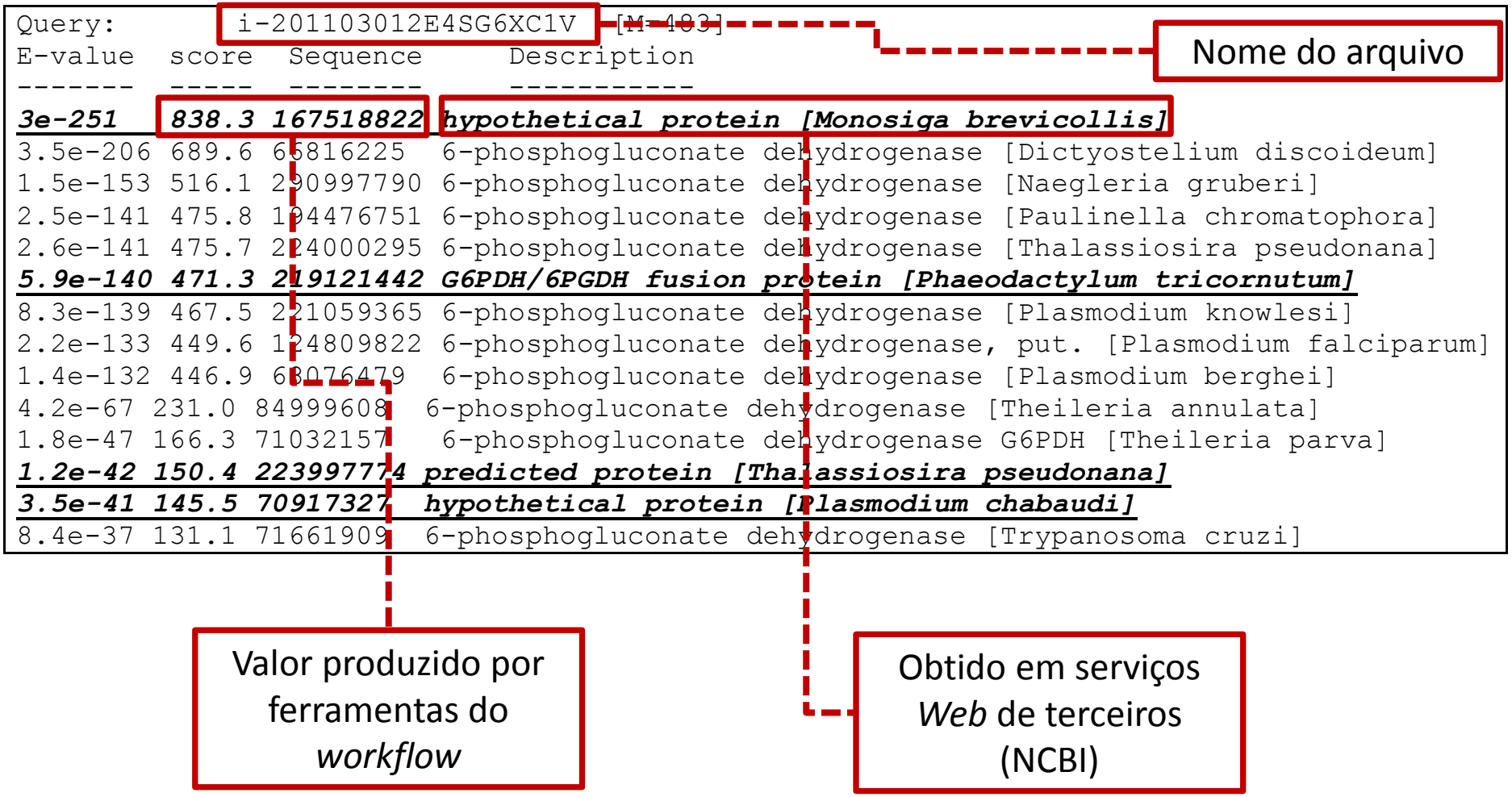
- Como extrair dados de diferentes formatos (e.g. binário, Fasta, HDF5, etc.) ?
- Componente extrator (baseado na álgebra)
  - Invoca um programa externo (definido pelo usuário) que analisa os arquivos produzidos e extrai os metadados de interesse
  - Encapsula regras de extração que são dependents de domínio
  - Relações possuem dados da execução com dados do domínio

Extração de dados de domínio

```
<SciCumulus>
  <database name="scicumulus" server="mp4-5b.dyndns.info" port="5432"/>
  <SciCumulusWorkflow tag="SciEvol" description="MER" exectag="scievol" expdir="/scievol/">
    <SciCumulusActivity tag="MSA" activation="./experiment.cmd" extractor="./extract_msa.cmd">
      <Relation reltype="Input" name="rel_in_1" filename="input_step_1.txt"/>
      <Relation reltype="Output" name="rel_out_1" filename="output_step_1.txt"/>
      <File filename="experiment.cmd" instrumented="true"/>
    </SciCumulusActivity>
  </SciCumulusWorkflow>
</SciCumulus>
```



Extração de Dados de Domínio



Consulta utilizando dados de domínio

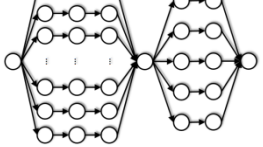
- Quais sequencias dadas como entrada não fazem parte de um gene específico?

```
SELECT S.NCBI_Ref_Sequence

FROM Task T, Hits H, MSA_CONVERTED M2, MSA M1,
                                SEQUENCE_GROUP SG,
SEQUENCE S,
                    ORGANISM O, SPECIE SP, GENUS G

WHERE T.taskid = H.taskid AND H.msacid = M2.msacid
AND M2.msaid = M1.msaid AND M1.sgid = SG.sgid
AND SG.sqid = S.sqid AND S.orgid = O.orgid
AND O.spid = SP.spid and SP.genid = G.genid
AND T.exitStatus = 0 /* No error */
AND G.genus = "PLASMODIUM"
```

Dados específicos do domínio

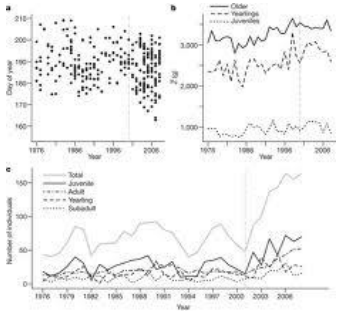
Roteiro do Tutorial

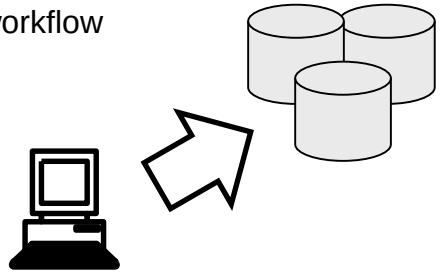
- Motivação
- Workflows Científicos
- Nuvens de Computadores
- Proveniência de Dados
- Máquina de Execução SciCumulus
- **Aplicação de Proveniência em e-Science na Nuvem**
- Demo



especificação do workflow

validação, verificação e publicação dos resultados





transferência de dados para a nuvem

composição

descoberta

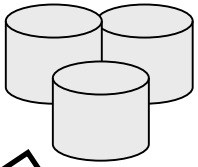
análise

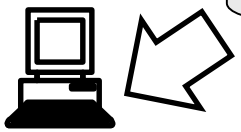
configuração

Dados de Proveniência

redimensionamento

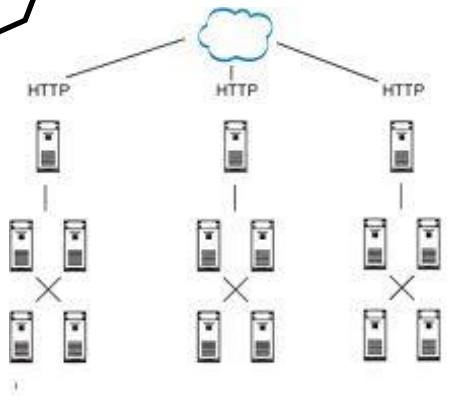
execução



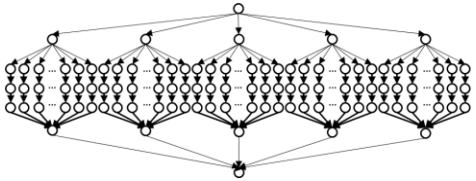


download de dados da nuvem

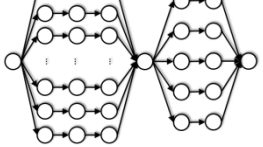
Configuração do cluster virtual



monitoramento da execução na nuvem



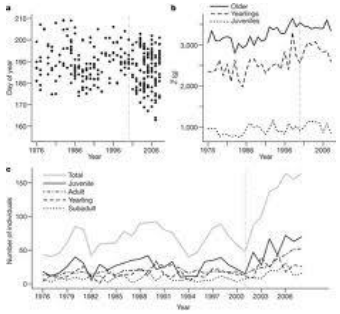
execução distribuída e paralela do workflow



especificação do workflow

Apoio nas diversas fases!

validação, verificação e publicação dos resultados



transferência de dados para a nuvem

configuração

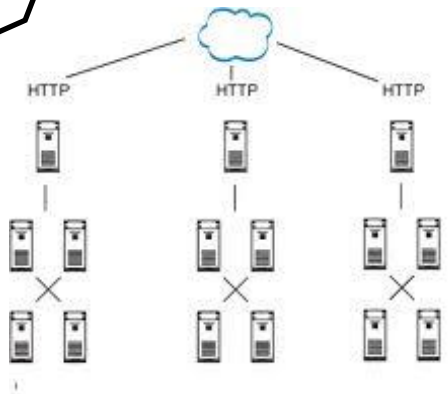
Dados de Proveniência

redimensionamento

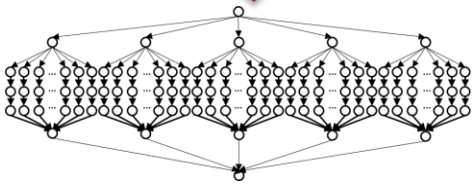
execução

download de dados da nuvem

Configuração do cluster virtual



monitoramento da execução na nuvem



execução distribuída e paralela do workflow

O uso da proveniência

Dimensionamento do Ambiente

- Recursos são instanciados sob demanda
- Vários tipos de máquinas virtuais passíveis de uso
  - Capacidade de armazenamento
  - CPU
  - Memória
  - Largura de banda
- Variação de provedor para provedor

Dimensionamento do Ambiente

- Recursos são instanciados sob demanda

Qual tipo utilizar?
Quantas máquinas instanciar ?
Força bruta ?

  - Memória
  - Largura de banda
- Variação de provedor para pro

Dimensionamento do Ambiente

Vários trabalhos existentes de dimensionamento de ambientes para aplicações comerciais

Dimensionamento do Ambiente

Dimensioning the Virtual Cluster for Parallel Scientific Workflows in Clouds

Daniel de Oliveira[1], Vitor Viana[3]
[1]IC/UFF
danielcmo@ic.uff.br

Eduardo Ogasawara[2]
[2]CEFET/RJ
eogasawara@cefet-rj.br

Kary Ocaña[3], Marta Mattoso[3]
[3]COPPE/UFRJ
{vviana,kary,marta}@cos.ufrj.br

ABSTRACT

Cloud computing has established itself as a solid computational model that allows for scientists to use a series of distributed virtual resources to execute a wide range of scientific [illegible] several cases, there is a demand [illegible] these [illegible] workflow in parallel may be a hard task to accomplish, *i.e.* it is hard to define and adapt the optimal number of virtual machines to be used. Most systems follow this manual configuration of the scientist for the whole workflow execution, using adaptive techniques only in the presence of failures. Due to the huge number of options (virtual machine types) to configure a cloud environment, the configuration task commonly becomes impractical to be performed

4th Workshop on Scientific Cloud Computing (ScienceCloud) 2013

Co-located with ACM HPDC 2013

New York City, NY, USA -- June 17th, 2013

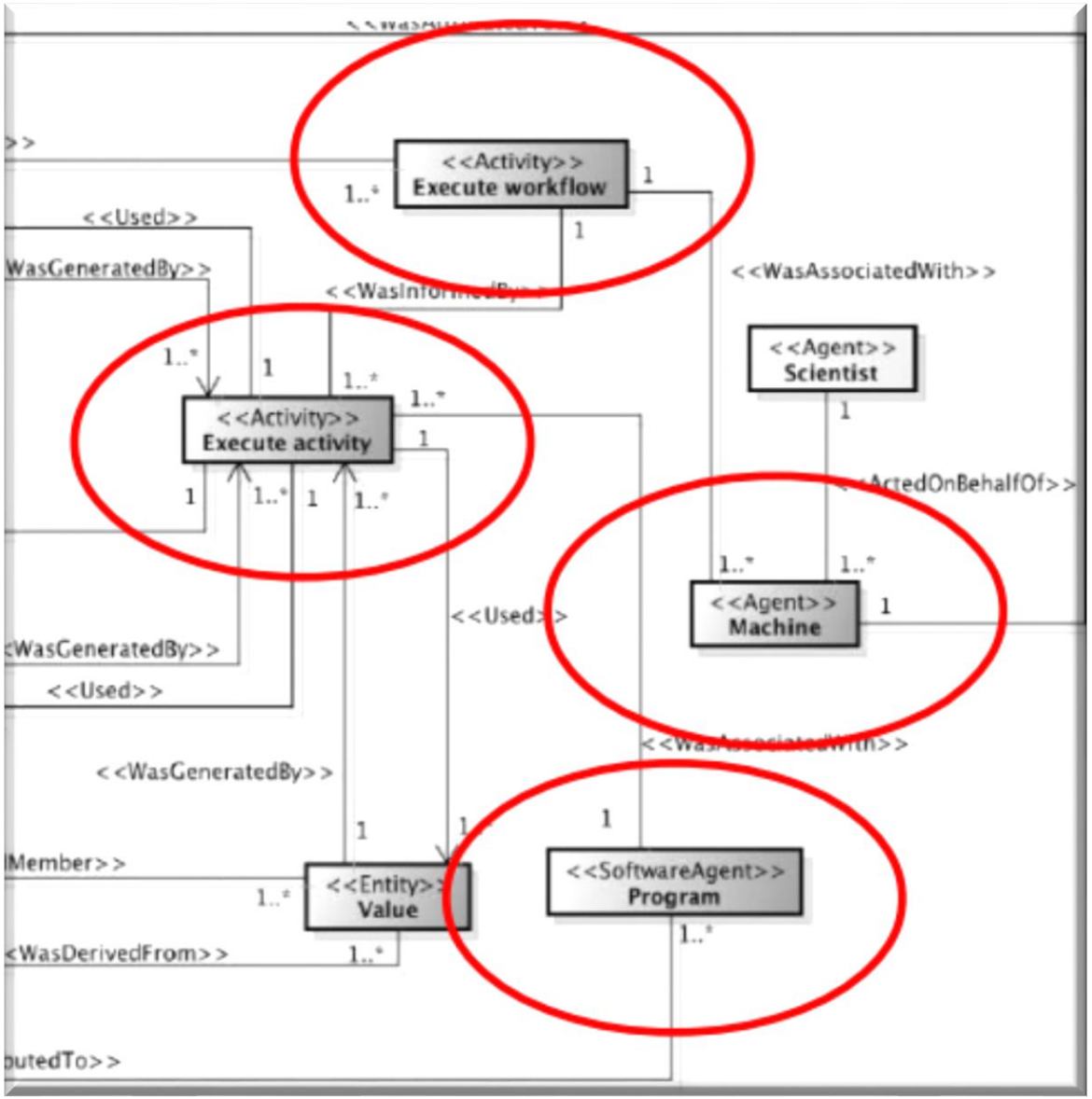
cloud computing for scientific workflows, [illegible] need HPC capabilities. In 2012 [illegible] simulation [illegible] to allocate virtual [illegible] expensive infrastructure elastically. [illegible] workflows tend to spend more time being analyzed than executed, generating idle computing time along the experiment life-cycle. This elastic allocation of resources, on demand, provides a new dimension for HPC applications.

Some Scientific Workflow Management Systems (SWfMS) [11,12] bridge the gap between cloud environments and the management of scientific experiments, such as, SciCumulus engine [4,13], Pegasus [14], Swift [15], and specifically for bioinformatics, the combination of Taverna and Galaxy named Tavaxy [16]. In these SWfMS a virtual cluster is created in the cloud to execute the

Dimensionamento do Ambiente

- Dimensionamento leva em consideração características do *workflow*
- Histórico de execução é utilizado
  - Proveniência retrospectiva
  - Características do ambiente
- Redução do tempo de execução e dos custos
  - Evita o super e o sub dimensionamento!

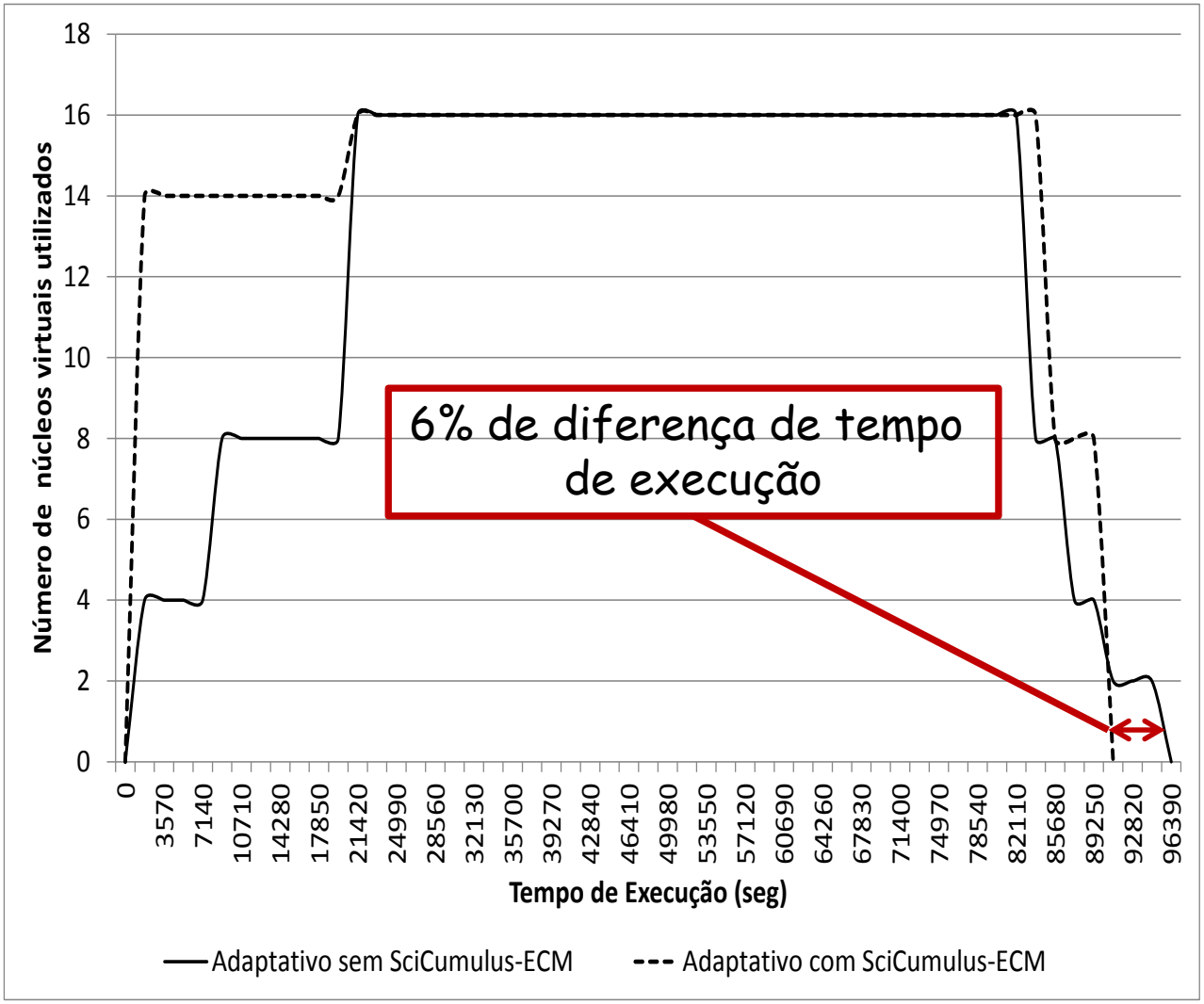
Quais dados são importantes?



SciCumulus-ECM

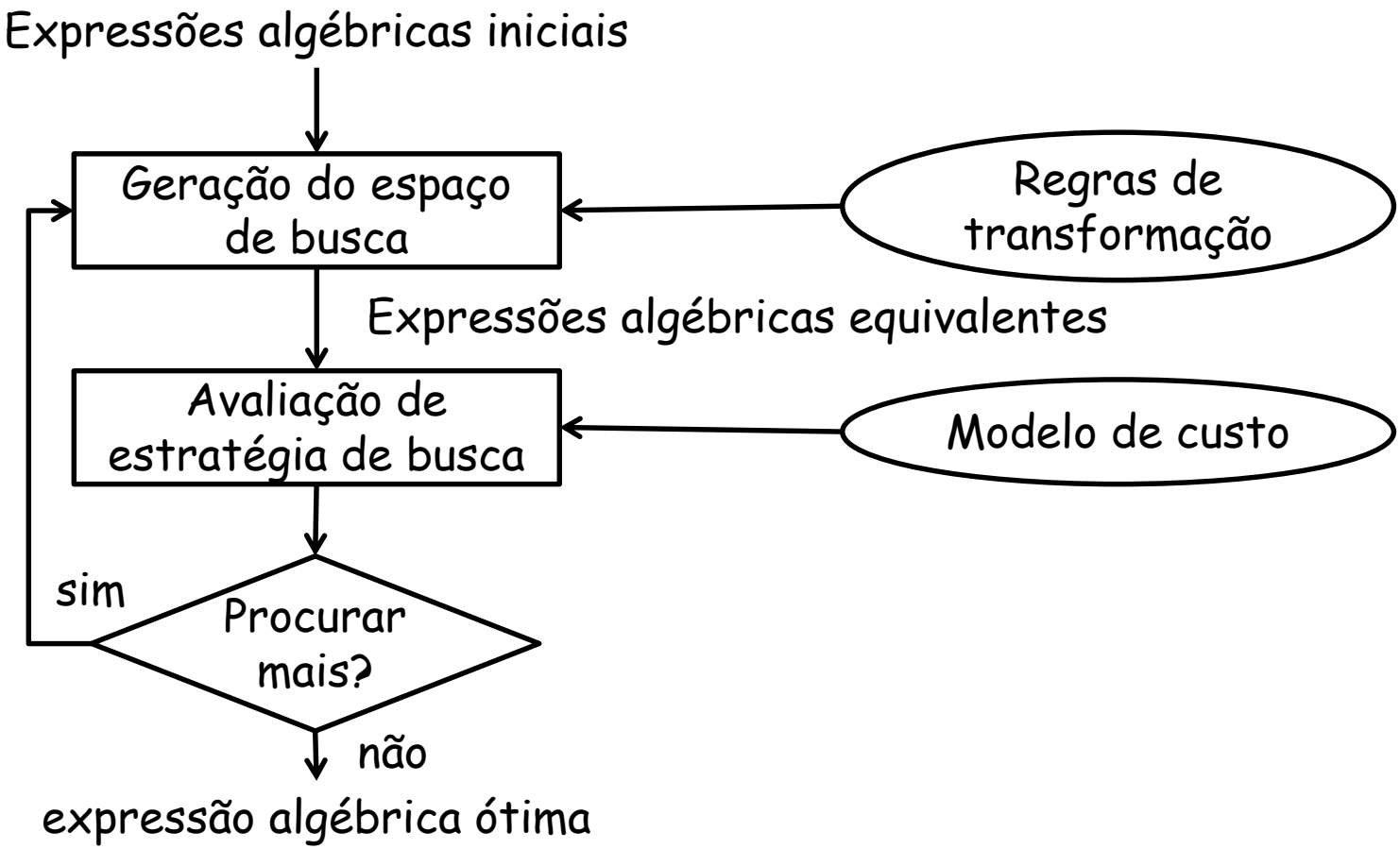
- Baseado em algoritmos genéticos
- Alimenta o algoritmo com os dados históricos de proveniência
- Com uma base de proveniência de tamanho médio (aprox. 100 execuções de *workflows*) o algoritmo apresenta uma convergência rápida
- Utilização do arcabouço JGAP

Análise da Execução Adaptativa Otimizando a Quantidade de Máquinas Virtuais



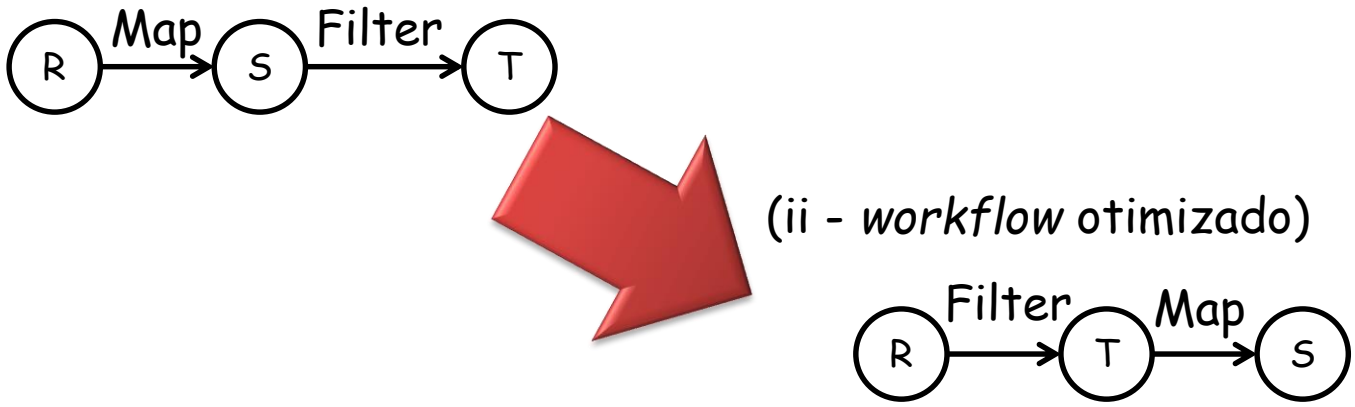
- 400 arquivos multi-fasta
- Cenário com foco no tempo de execução
- Orçamento limite de US$ 75,00

Processo de otimização de workflows



Transformações algébricas

(i - *workflow* - perspectiva de relações)



- Garantia do esquema das atividades
- Similar a otimização de consultas na álgebra relacional
- Antecipação do Filtro $\cong$ Antecipação da seleção

Escalonamento Adaptativo

- O ambiente de nuvem possui instabilidades
  - Falhas nas VMs
  - Aplicação de atualizações
  - Reinicializações
  - Heterogeneidade das VMs
- O escalonamento deve levar essas características em consideração

Escalonamento Adaptativo

– $f(a_i, VM_j) =$

$\alpha_1 \times$(*Tempo de Execução + Tempo de Transferência*)
+
$\alpha_2 \times$ *Custo de Confiabilidade*
+
$\alpha_3 \times$(*Vh* × [*Tempo de Execução + Tempo de Transferência*] + *Vt* ×*Quantidade de Dados Transferidos*)

- *A* a_i escolhida é aquela que satisfaz $f(a_i, VM_j) = min \ \forall \ VM_j \in VM \ \{f(a_i, VM_j)\}$

Escalonamento Adaptativo

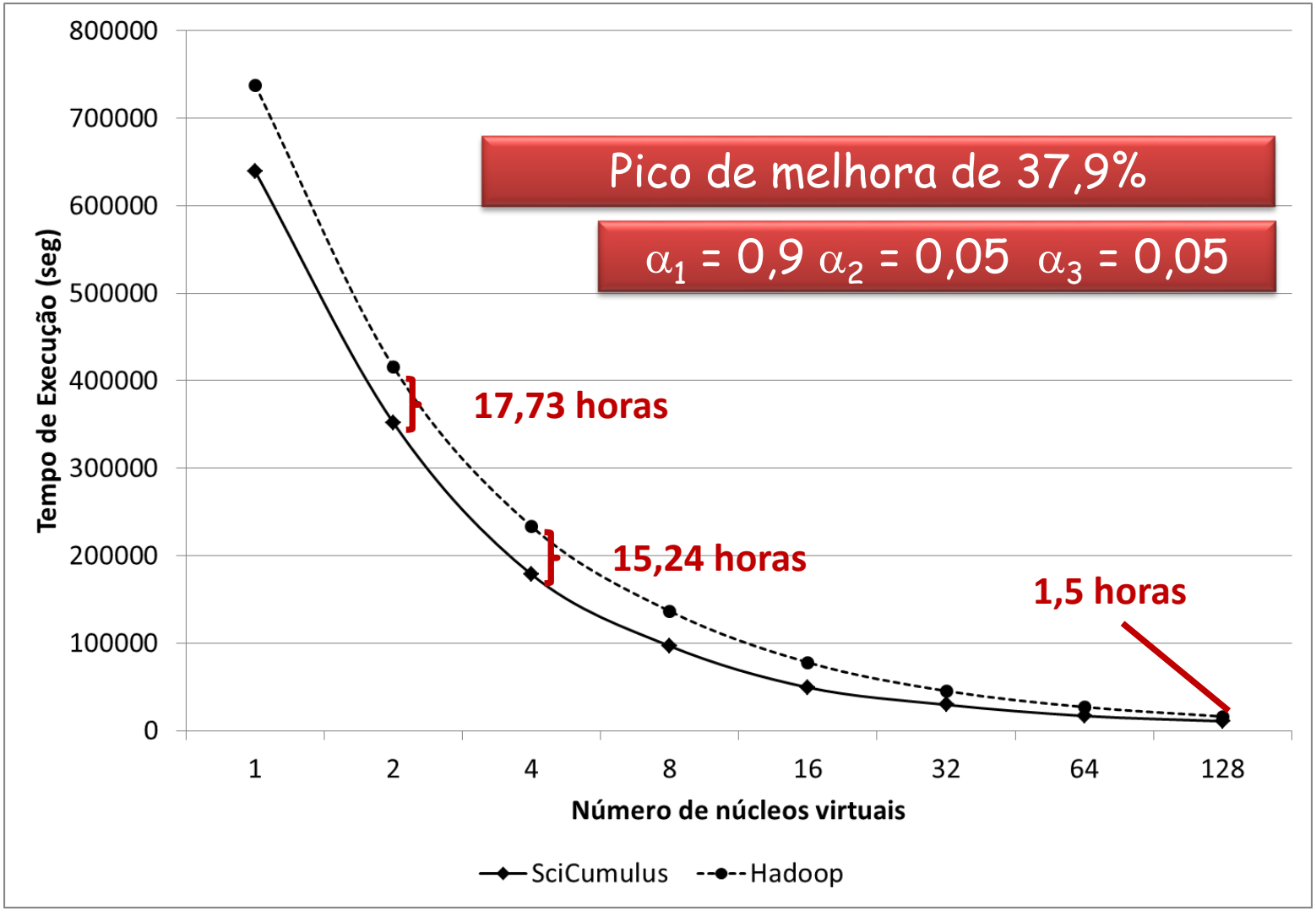
– $f(a_i, VM_j)$ =

α_1×(Tempo de Execução
+
α_2× Custo de C
+
α_3×(

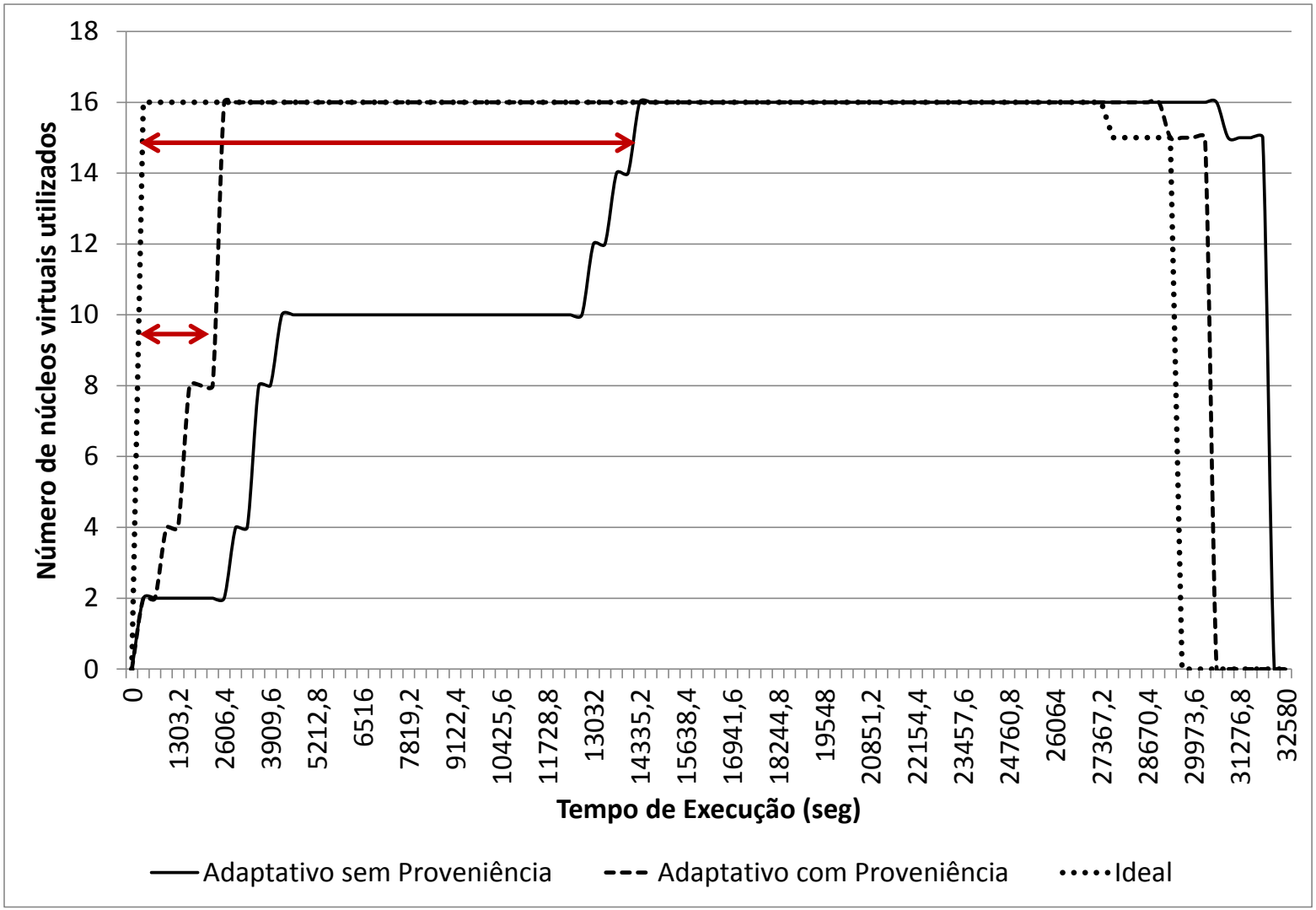
Todas as informações são buscadas no repositório de proveniência e levam em consideração a operação algébrica

- que satisfaz $f(a_i, VM_j)$ = $\{f(a_i, VM_j)\}$

Execução Adaptativa – sem variação na quantidade de VMs

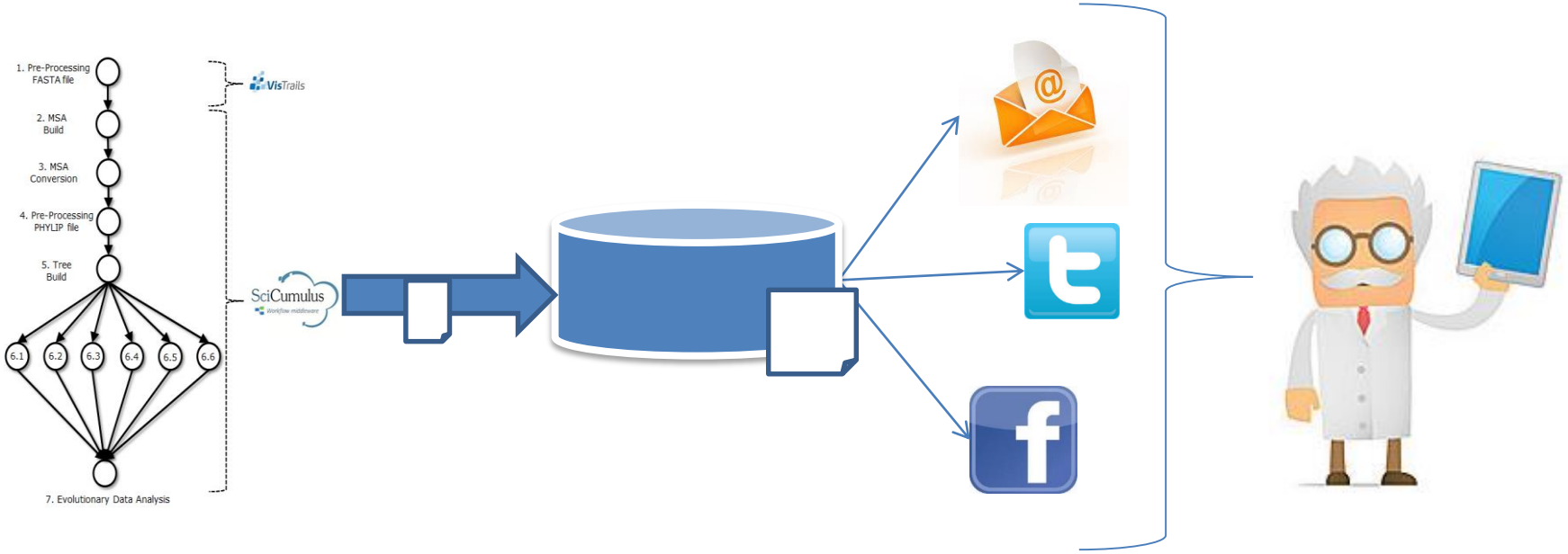


Execução Adaptativa – variação na quantidade de VMs



Monitoramento

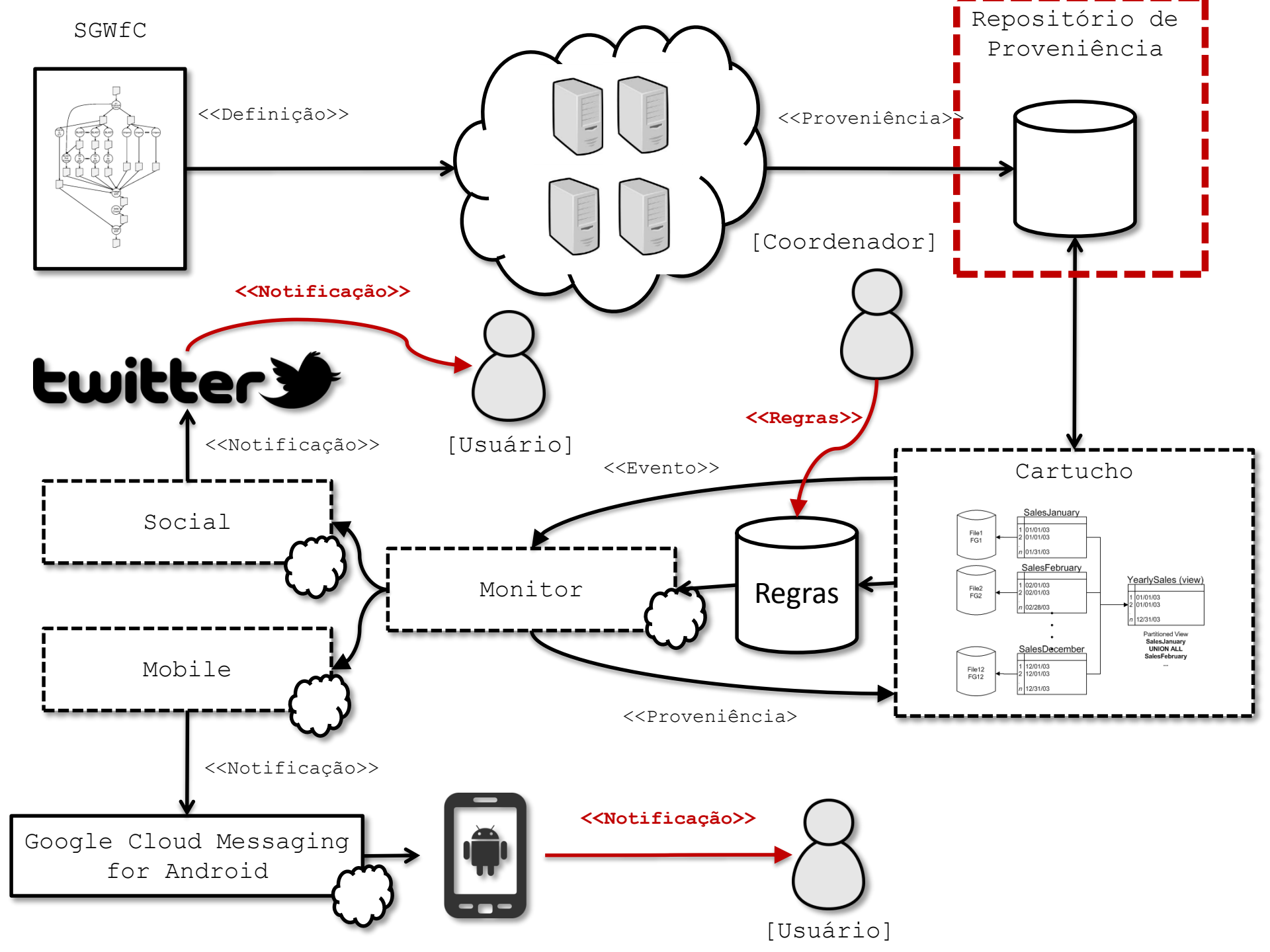
- Execuções podem durar semanas ou meses
- Monitoramento *in situ* é inviável
- Utilização de redes sociais e computação móvel



SciLightning

- Serviço de monitoramento de execução de workflows científicos em paralelo
- 3 componentes principais:
  - SciLightning Monitor
  - SciLightning Social
  - SciLightning Mobile
- Todo o monitoramento é baseado nos dados de proveniência gerados pelo SciCumulus

PINTAS, J. ; OLIVEIRA, D. ; OCANA, K. ; DIAS, J. ; MATTOSO, M. L. Q. . **Monitoramento em Tempo Real de Workflows Científicos Executados em Paralelo em Ambientes Distribuídos**. In: VI e-Science Workshop (em conjunto com CSBC 2012), 2012, Curitiba. Anais do VI e-Science Workshop (em conjunto com CSBC 2012). Porto Alegre : Sociedade Brasileira de Computação, 2012.


SGWfC
<<Definição>>
<<Proveniência>>
Repositório de Proveniência
[Coordenador]
<<Notificação>>
twitter
[Usuário]
<<Regras>>
<<Notificação>>
<<Evento>>
Cartucho
Social
Monitor
Regras
Mobile
SalesJanuary
SalesFebruary
SalesDecember
YearlySales (view)
Partitioned View
SalesJanuary
UNION ALL
SalesFebruary
<<Proveniência>>
<<Notificação>>
Google Cloud Messaging for Android
<<Notificação>>
[Usuário]

SciLightning

Recuperação de Falhas

- Falha de VMs é um problema constante
- O provedor garante a reposição de uma VM, mas não de seu contexto
- O cientista não deve re-executar atividades que já foram executadas anteriormente
- Proveniência fornece esse tipo de informação de contexto da execução

SciMultaneous

- Serviço de re-execução de *atividades*
- Explora características da nuvem como elasticidade por meio de duas heurísticas
  - H1: **Redundância de atividades**
  - H2: **Monitoramento contínuo**

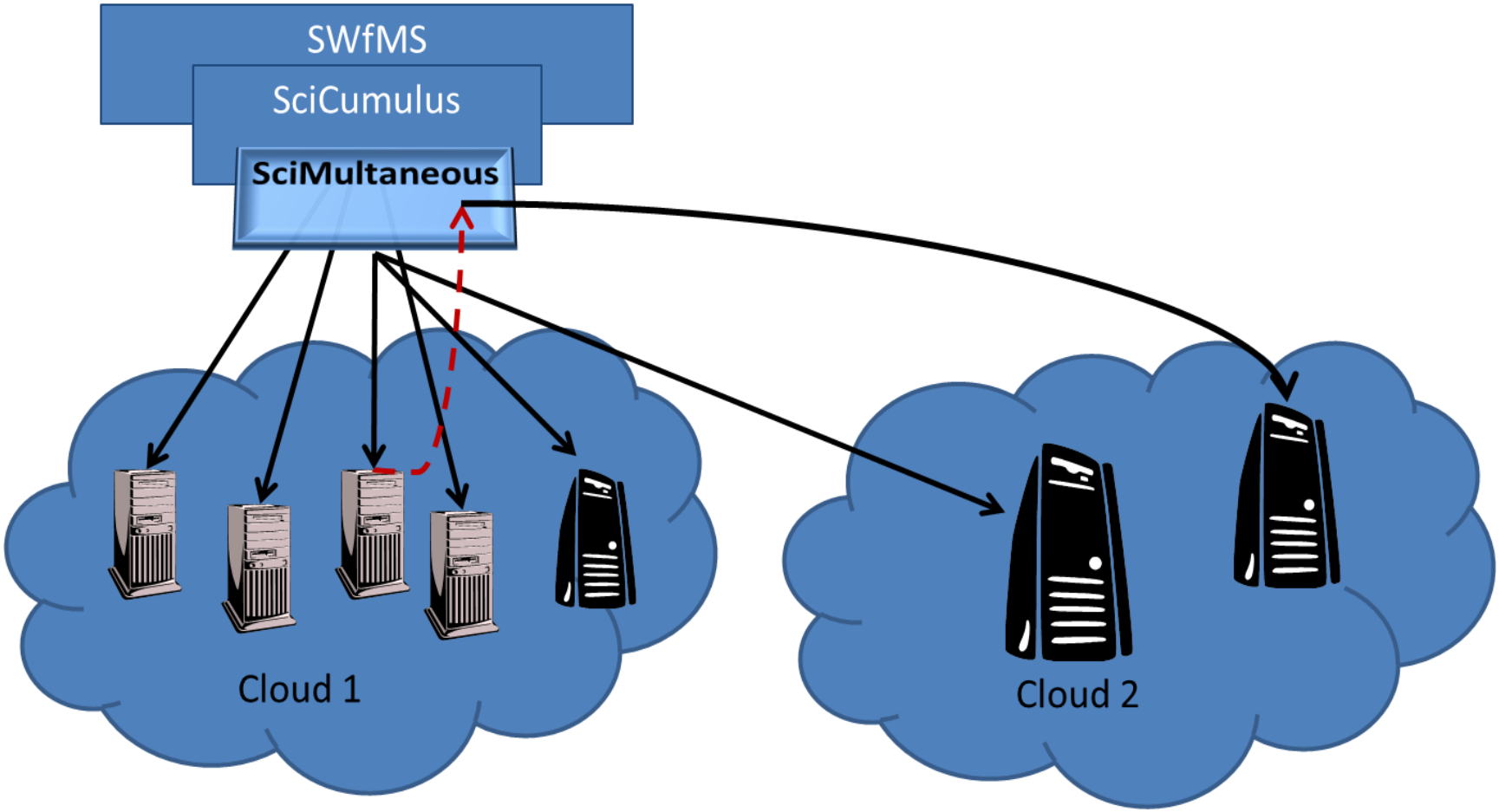
COSTA, F. ; OLIVEIRA, D. ; OCANA, K. ; OGASAWARA, E. ; MATTOSO, M. L. Q. . **Enabling Re-Executions of Parallel Scientific Workflows Using Runtime Provenance Data**. In: 4th International Provenance and Annotation Workshop, 2012, Santa Barbara, CA. Proceedings of the 4th International Provenance and Annotation Workshop. Heildelberg : Springer Verlag, 2012.

Consulta de Proveniência

- "Recuperar, por ordem crescente de execuções dos *workflows*, as datas de início e término, *tags* dos *workflows*, bem como o nome de todas as atividades associadas e que contenham alguma *execução* com erro".

```
SELECT w.tag,
       a.tag,
       t.exitstatus,
       t.processor,
       t.workspace,
       t.status,
       t.endtime,
       t.starttime,
       extract ('epoch' from (t.endtime-t.starttime))||','  as duration
from h_workflow_ w, h_activity_ a, hactivation t
where w.wkfid = a.wkfid
and a.actid = t.actid
and not exists (select * from hactivation a2
                where a2.actid = a.actid
                and a2.exitstatus <> 0)
order by w.wkfid
```

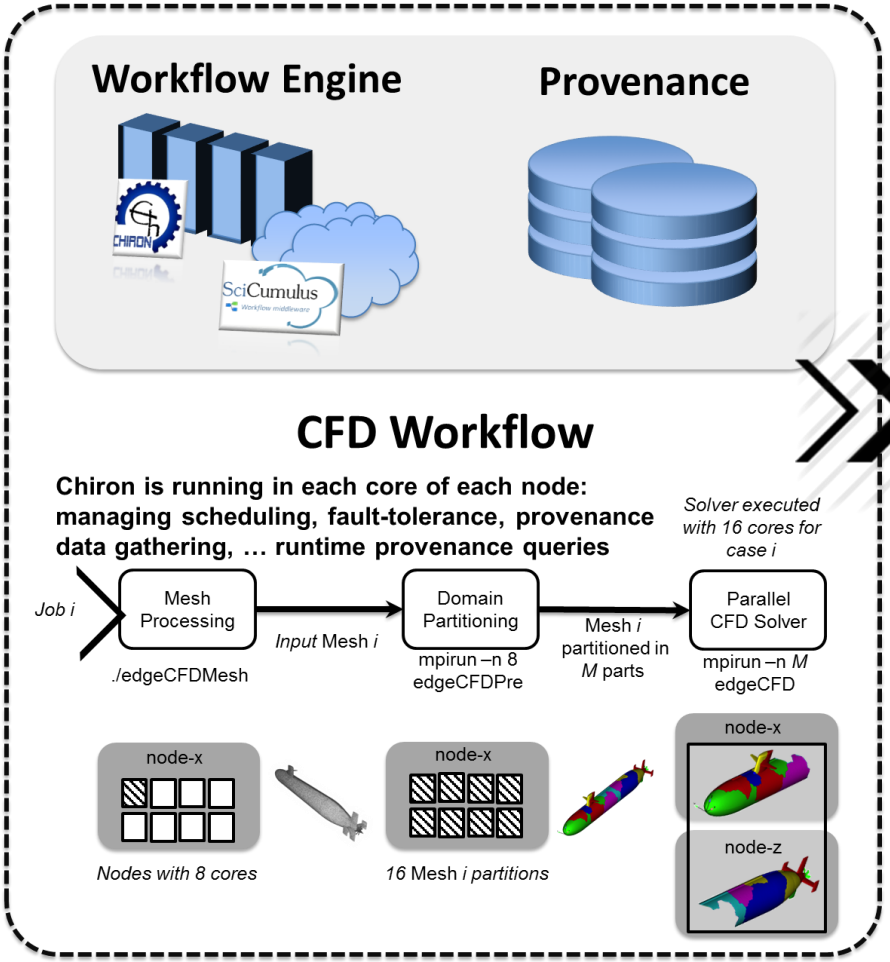
SciMultaneous



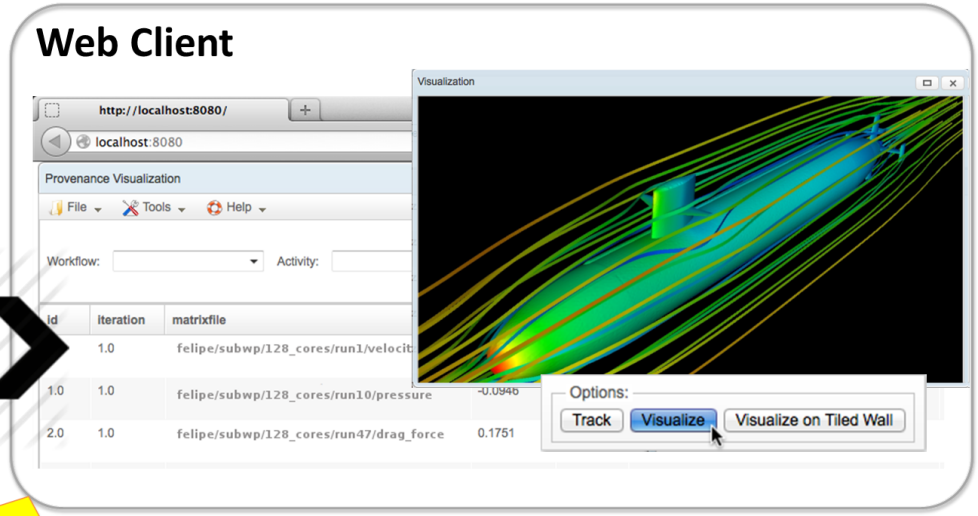
Visualização

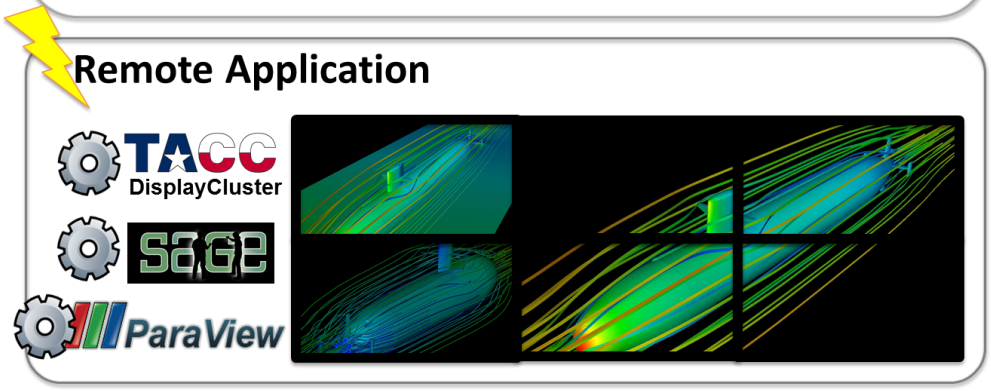
- A visualização dos dados dos experimentos é uma tarefa fundamental de análise
- Grande volume de dados
- Necessidade de associar metadados a imagens e vídeos produzidos
- Necessidade de utilização de ambientes de visualização como *tiled walls displays*

PROV-Vis



Prov-Vis





PROV-Vis


Web Client
http://localhost:8080/
localhost:8080
Provenance Visualization
File
Tools
Help
Workflow:
Activity:
id
iteration
matrixfile
0.0
1.0
1.0
1.0
2.0
1.0
felipe/subwp/128_cores/run47/drag_force
0.1751
Visualization
Dados de Proveniência
Track
Visualize
Visualize on Tiled Wall
Remote Application

Roteiro do Tutorial

- Motivação
- Workflows Científicos
- Nuvens de Computadores
- Proveniência de Dados
- Máquina de Execução SciCumulus
- Aplicação de Proveniência em e-Science
- **Demo**

Demo SciCumulus

- Utilização do ambiente Amazon EC2
  - Necessário que o usuário possua uma conta
- Imagem pública: ami-6e1a8907
- Linux CentOS 5.5
- SciCumulus
  - java -jar scicumulus-core.jar 0 /root/expSciPhy/machines.conf niodev MPI 2 /root/expSciPhy/scicumulus.xml

Download

Home / Browse / Science & Engineering / Bio-Informatics / SciCumulus / Support

Prealpha

Brought to you by: eogasawara, jonasdias, phydoop

Looking for the latest version? **Download SciCumulusCore-0.1-SNAPSHOT.jar (6.8 MB)**

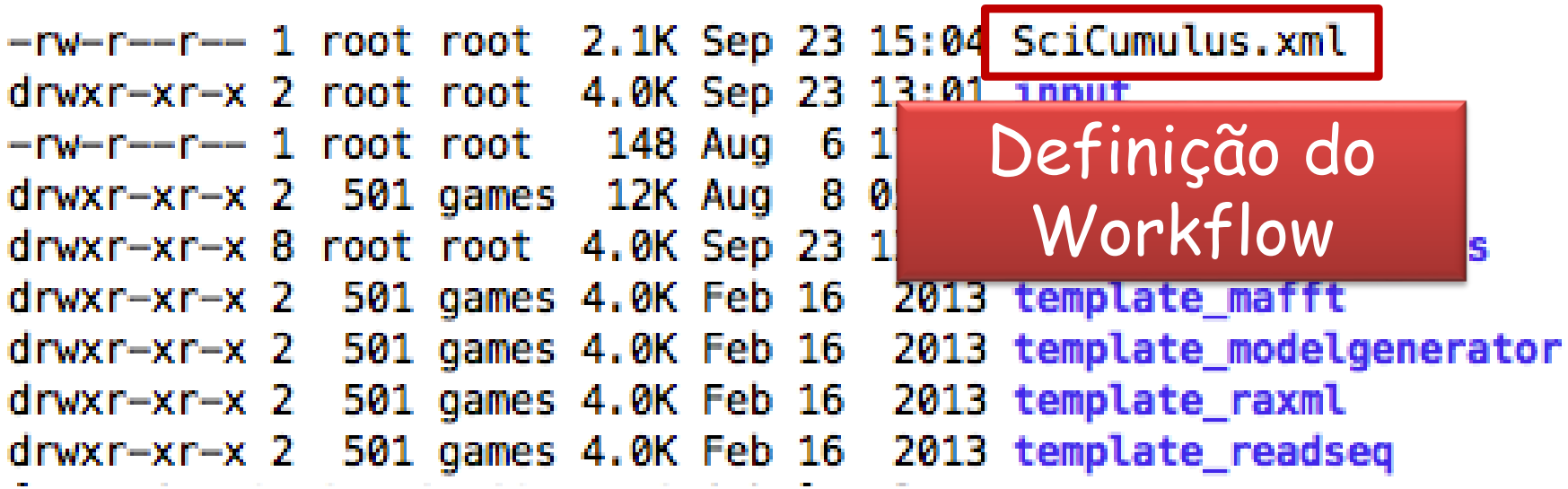
Home

Name	Modified	Size	Downloads
SciCumulus_compact.backup	2013-03-20	11.6 MB	0
SciCumulusCore-0.1-SNAPSHOT.jar	2013-03-20	6.8 MB	2
SciCumulusDesktop-1.0-SNAPSHOT.jar	2013-03-20	18.3 kB	1
sciCumulus.xml	2013-03-20	687 Bytes	1

Configuração do Workflow

```
-rw-r--r-- 1 root root  2.1K Sep 23 15:04 SciCumulus.xml
drwxr-xr-x 2 root root  4.0K Sep 23 13:01 input
-rw-r--r-- 1 root root   148 Aug  6 17:10 machines.conf
drwxr-xr-x 2  501 games  12K Aug  8 05:00 output
drwxr-xr-x 8 root root  4.0K Sep 23 13:10 result_bioscicumulus
drwxr-xr-x 2  501 games 4.0K Feb 16  2013 template_mafft
drwxr-xr-x 2  501 games 4.0K Feb 16  2013 template_modelgenerator
drwxr-xr-x 2  501 games 4.0K Feb 16  2013 template_raxml
drwxr-xr-x 2  501 games 4.0K Feb 16  2013 template_readseq
```

Configuração do Workflow



Configuração do Workflow

```
-rw-r--r-- 1 root root  2.1K Sep 23 15:04 SciCumulus.xml
drwxr-xr-x 2 root root  4.0K Sep 23 13:01 input
-rw-r--r-- 1 root root   148 Aug  6 17:10 machines.conf
drwxr-xr-x 2  501 games  12K Aug  8 05:00 output
drwxr-xr-x 8 root root  4.0K Sep 23 13:10 result_bioscicumulus
drwxr-xr-x 2  501 games 4.0K Feb 16  2013 template_mafft
drwxr-xr-x 2  501 games 4.0K Feb 16  2013 template_modelgenerator
drwxr-xr-x 2  501 games 4.0K Feb 16  2013 template_phyml
drwxr-xr-x 2  501 games 4.0K Feb 16  2013 template_raxml
```

Definição dos templates das atividades

Configuração do Workflow

```
<Hydra>
    <database name="scicumulus" password="livre2008!" port="5432" server="ec2-50-17-107-164.compute-1.amazonaws.com" username="pos
    <HydraWorkflow description="Phylogenetic tree construction" exectag="exsciphy-2" expdir="/root/exp_SciPhy" tag="SciPhy">

        <HydraActivity activation="./experiment.cmd" description="alignment" tag="mafft" templatedir="/root/exp_SciPhy/template_ma
                <Relation filename="parameter.txt" name="A" reltype="Input" />
                <Relation filename="output_mafft.txt" name="C" reltype="Output" />
                <File filename="experiment.cmd" instrumented="true" />
                <Field input="A" name="FASTA_FILE" output="C" type="string" />
        </HydraActivity>
        <HydraActivity activation="./experiment.cmd" description="Format Alignment" tag="readseq" templatedir="/root/exp_SciPhy/te
                <Relation dependency="mafft" filename="output_mafft.txt" name="A" reltype="Input" />
                <Relation filename="output_readseq.txt" name="C" reltype="Output" />
                <File filename="experiment.cmd" instrumented="true" />
                <Field input="A" name="FASTA_FILE" output="C" type="string" />
        </HydraActivity>
        <HydraActivity activation="./experiment.cmd" description="Search for models" tag="modelgenerator" templatedir="/root/exp_S
                <Relation dependency="readseq" filename="output_readseq.txt" name="A" reltype="Input" />
                <Relation filename="output_modelgenerator.txt" name="C" reltype="Output" />
                <File filename="experiment.cmd" instrumented="true" />
                <Field input="A" name="FASTA_FILE" output="C" type="string" />
        </HydraActivity>
        <HydraActivity activation="./experiment.cmd" description="Phylogeny with RAxML" tag="raxml" templatedir="/root/exp_SciPhy/
                <Relation dependency="modelgenerator" filename="output_modelgenerator.txt" name="A" reltype="Input" />
                <Relation filename="output_raxml.txt" name="C" reltype="Output" />
                <File filename="experiment.cmd" instrumented="true" />
                <Field input="A" name="FASTA_FILE" output="C" type="string" />
        </HydraActivity>
        </HydraWorkflow>
</Hydra>
```

Representação dos *Workflows*

```
<?xml version="1.0" standalone="no"?>
<SciCumulus>
   <database name="scicumulus_adaptive" server="mp4-4.dyndns.info"
      port="5432" username="scicumulus" password="**********" />
   <SciCumulusWorkflow tag="filogenia" description="This is a test using anflex."
      exectag="Experimento Kary/200 organismos - Adaptive" expdir="/root/exp">
      <SciCumulusActivity tag="mafft" description="mafft" type="MAP"
            templatedir="/root/exp/template_mafft" activation="experiment.cmd">
         <Relation reltype="Input" name="A" filename="parameter.txt" />
         <Field name="NAME" type="string" />
         <Field name="FASTA_FILE" type="string" />
         <Relation reltype="Output" name="C" filename="output_mafft.txt" />
         <Field name="NAME" type="str
         <Field name="NUM_SEQ" type=
         <Field name="FASTA_FILE" ty
         <File filename="experiment.cmd
      </SciCumulusActivity>
   </SciCumulusWorkflow>
</SciCumulus>
```

```
NAME;FASTA_FILE
G1;ORTHOMCL2033
G2;ORTHOMCL1895
G3;ORTHOMCL2034
G4;ORTHOMCL1896
G5;ORTHOMCL2035
G6;ORTHOMCL1897
G7;ORTHOMCL2036
G8;ORTHOMCL1898
G9;ORTHOMCL2037
G10;ORTHOMCL1899
```

Configuração do Workflow

```
# Number of Processes
1
# Protocol switch limit
131072
# Entry in the form of machinename@port@rank
ec2-54-242-8-34.compute-1.amazonaws.com@22000@0
```

Configuração do Workflow

```
drwxr-xr-x 2 root root 4.0K Sep 23 13:10 1
drwxr-xr-x 2 root root 4.0K Sep 23 13:10 2
drwxr-xr-x 2 root root 4.0K Sep 23 13:10 3
drwxr-xr-x 2 root root 4.0K Sep 23 13:10 4
drwxr-xr-x 2 root root 4.0K Sep 23 13:10 5
```

Configuração do Workflow

```
-rw-r--r-- 1 root root 2.6K Sep 23 13:10 ENZ1.fasta
-rw-r--r-- 1 root root 1.4K Sep 23 13:10 ENZ1.fastaNumbered
-rw-r--r-- 1 root root 1.5K Sep 23 13:10 ENZ1.mafft
-rw-r--r-- 1 root root 1.3K Sep 23 13:10 H_Err.txt
-rw-r--r-- 1 root root   24 Sep 23 13:10 H_Relation.txt
-rw-r--r-- 1 root root    0 Sep 23 13:10 H_Result.txt
-rwxr-xr-x 1 root root  383 Sep 23 13:10 experiment.cmd
-rwxr-xr-x 1 root root 1.2K Sep 23 13:10 numberFasta.pl
```

Quer saber mais sobre banco de dados e nuvens?

- **Não perca!**
- "Análise em Big Data e um Estudo de Caso utilizando Ambientes de Computação em Nuvem"
  - Ticiana L. C. da Silva, Flávio R. C. Sousa, José Antônio F. de Macêdo e Javam C. Machado
  - Quinta-feira 03 de outubro – 09:00 – Sala Imperial

Agradecimentos

SBBD
28º Simpósio Brasileiro de Banco de Dados 2013

Obrigado!

Daniel de Oliveira – danielcmo@ic.uff.br
Marta Mattoso – marta@cos.ufrj.br

Instituto de computação

COPPE UFRJ