SBBD SBES 2009

V WAAMD

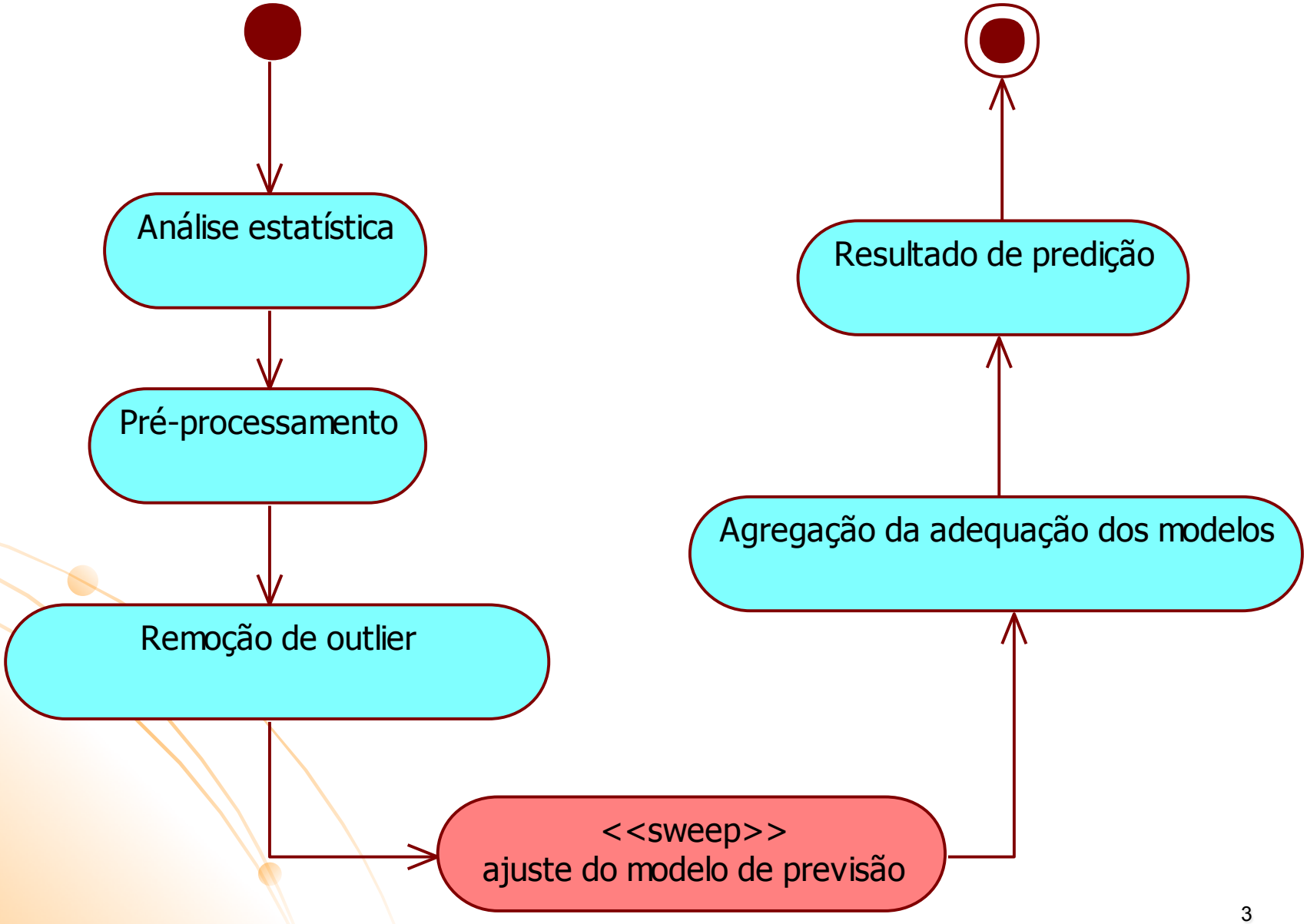
Paralelização de Tarefas de Mineração de Dados Utilizando Workflows Científicos

Carlos Barbosa
Daniel de Oliveira
Eduardo Ogasawara
Marta Mattoso

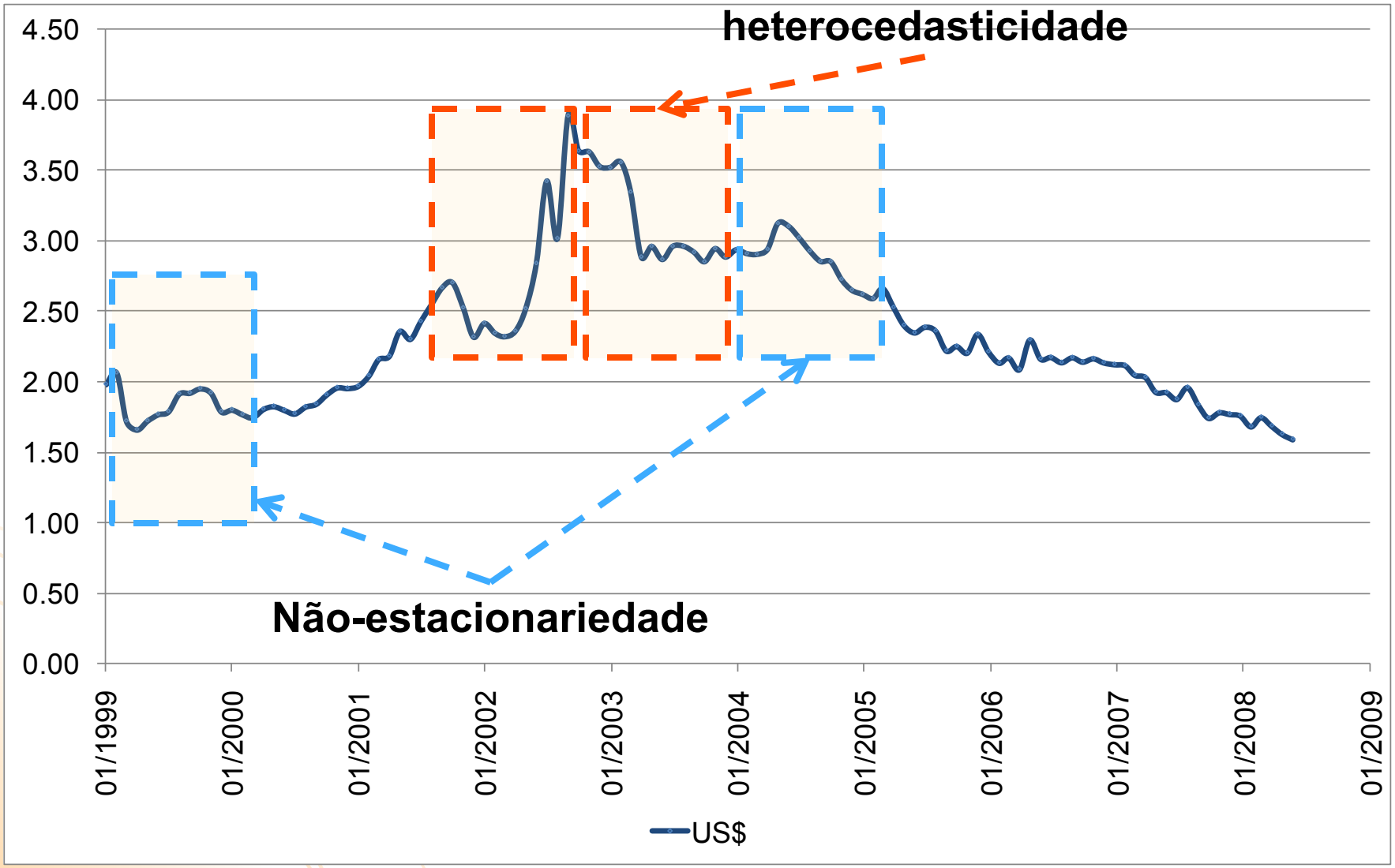
Sumário

- Processo de mineração de dados para séries temporais (ST)
- Utilização de workflows científicos para mineração de dados
- Paralelização do processo de mineração
- Resultados experimentais
- Trabalhos relacionados
- Conclusões

Processo de mineração de ST



Análise estatística



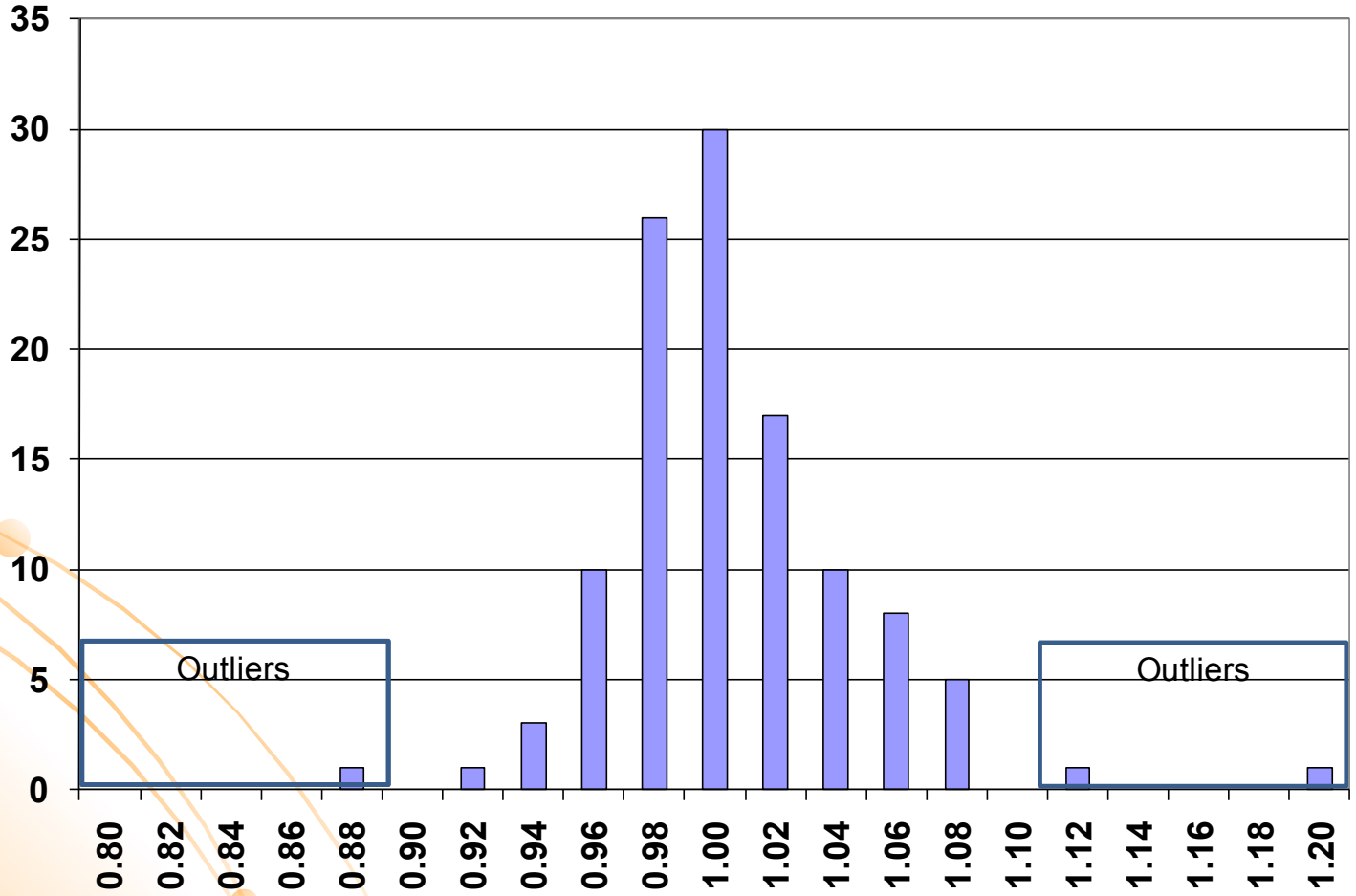
Estacionariedade: Teste de raiz unitário ou Teste Dickey-Fuller

Pré-processamento

- Formas de coleta
  - Global
  - Sliding windows
- Transformações
  - Remoção de tendência
  - Diferenciação
  - Log/Return
  - Normalização adaptativa*

Remoção de *outliers*

Técnicas lineares e por distribuição



Modelo de previsão

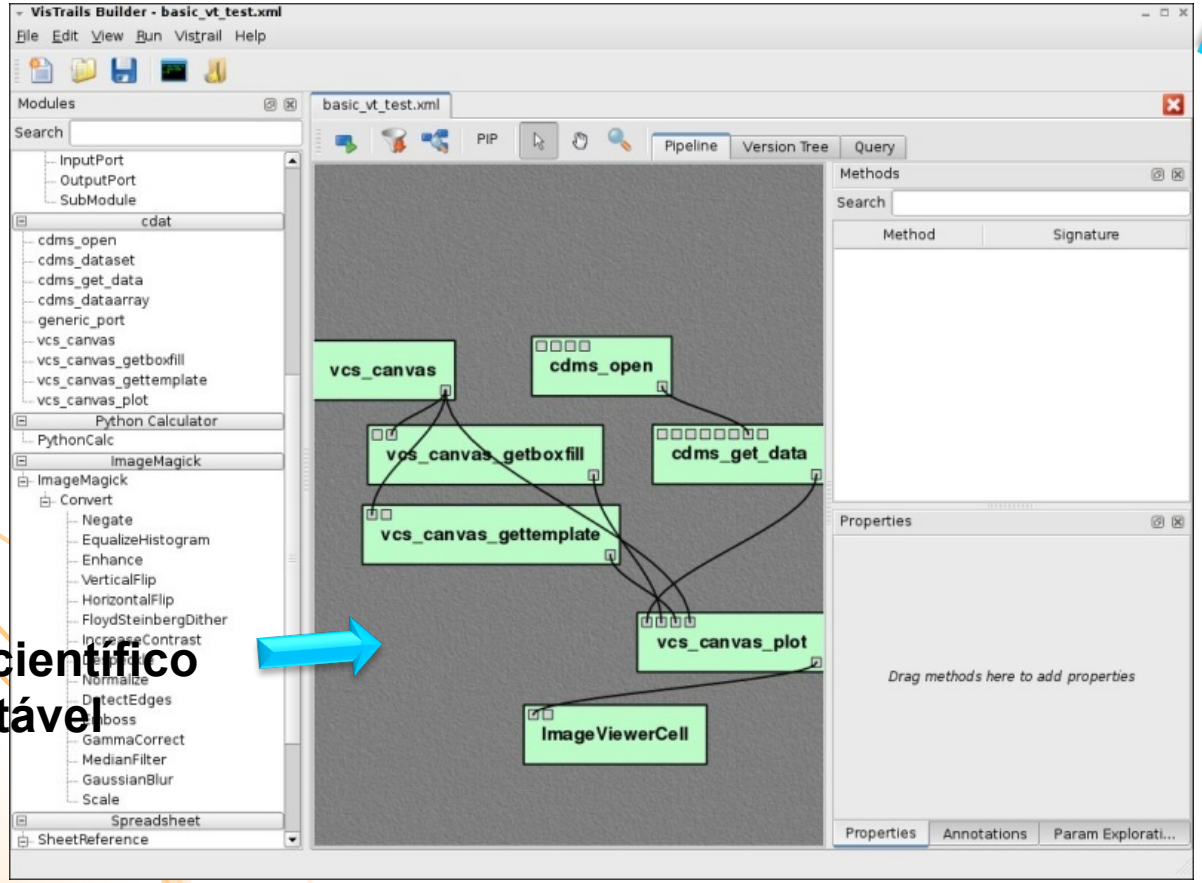
- Técnicas mais usadas:
  - Regressões lineares
- Outras técnicas lineares:
  - Auto regressão (*Box-Jekins*)
  - Vetores Auto-Regressivos
- Técnicas não-lineares:
  - Máquinas de vetores de suporte (SVM)
  - Redes Neurais (RN)*

Processo de Mineração

- Muitas alternativas ...
- Como foi resolvida a estacionariedade?
- Qual foi o método usado?
- Quais foram os parâmetros usados?
- Onde estão os dados, arquivos e programas que deram o melhor resultado?
- Qual foi mesmo o critério usado para identificar a melhor configuração?

Uso de Workflows Científicos

- Representa o fluxo de atividades como uma tentativa de solução para o experimento



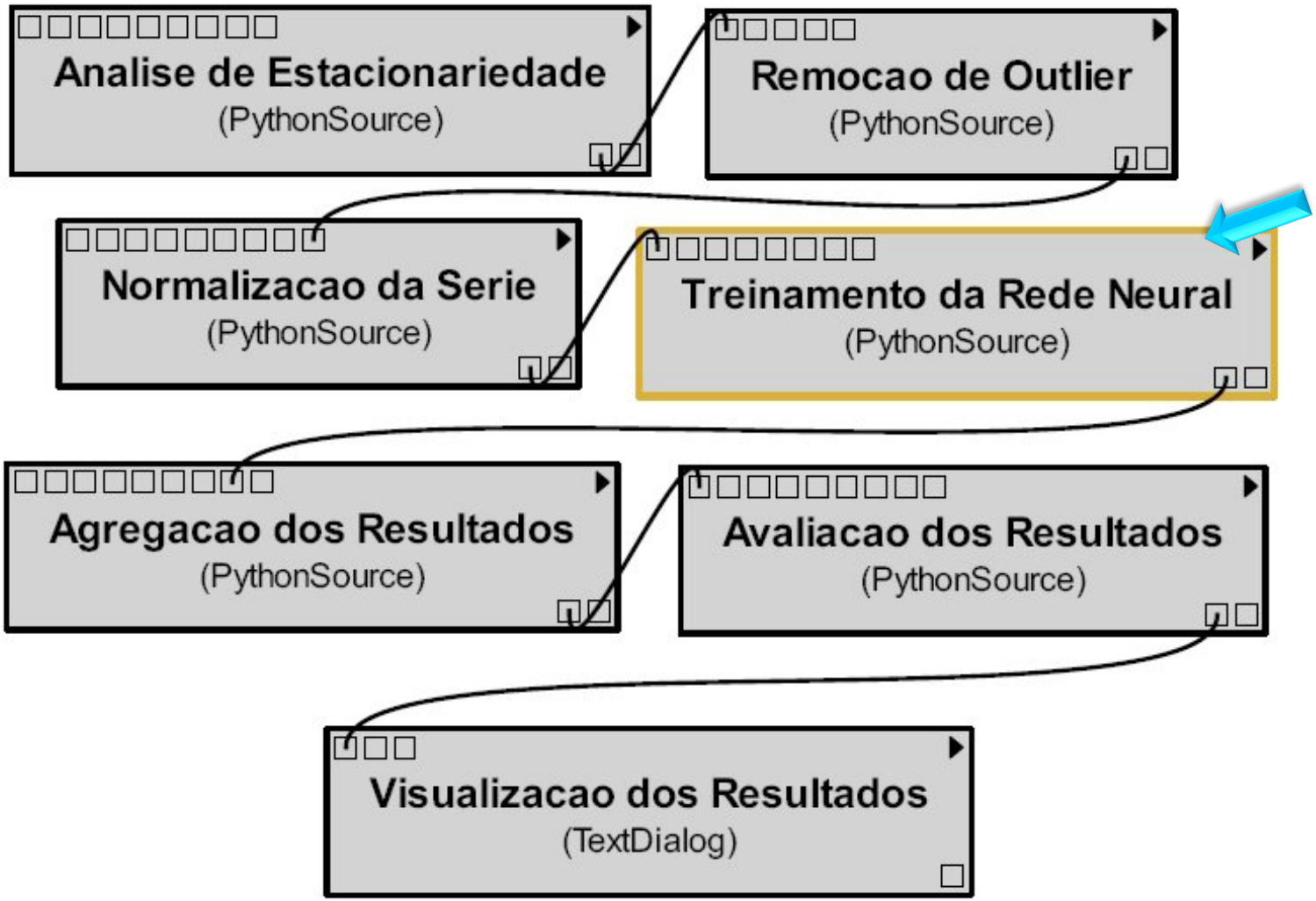
SGWfC

Workflow científico executável

Características dos Workflows Científicos

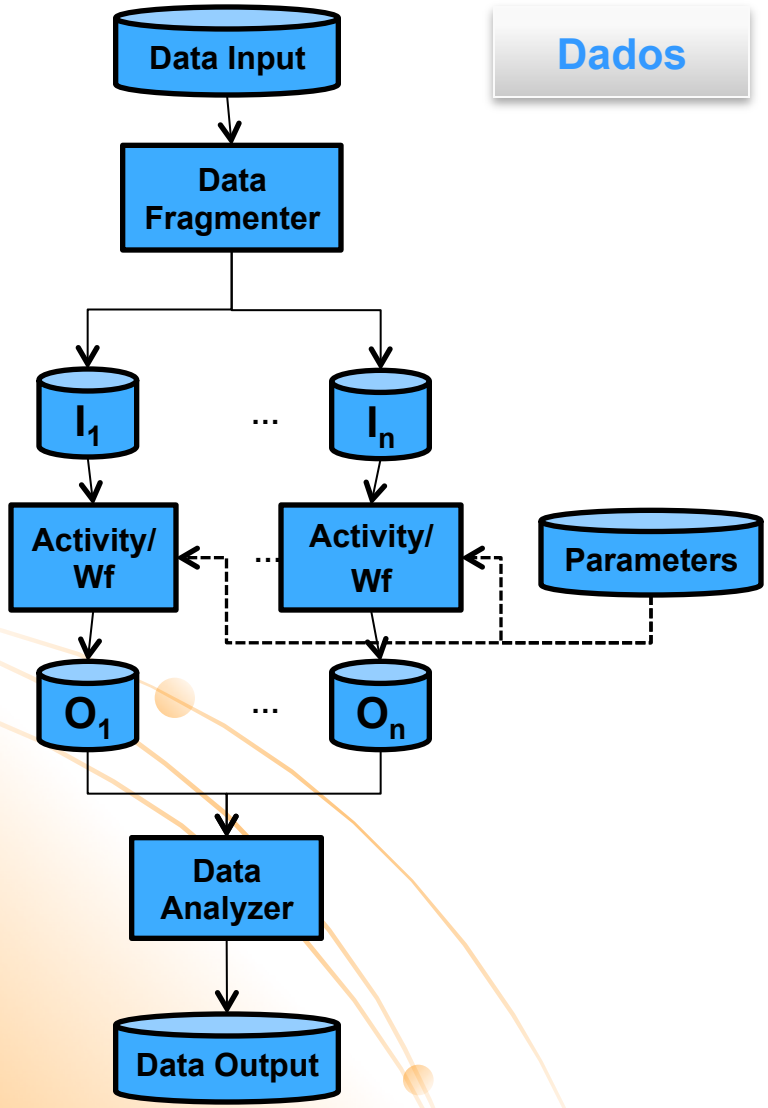
- Documentação do processo
- Apoio à proveniência
- Exploração de parâmetros
- Comparação visual
- Gerência de configuração
- Primitivas para manipulação de dados

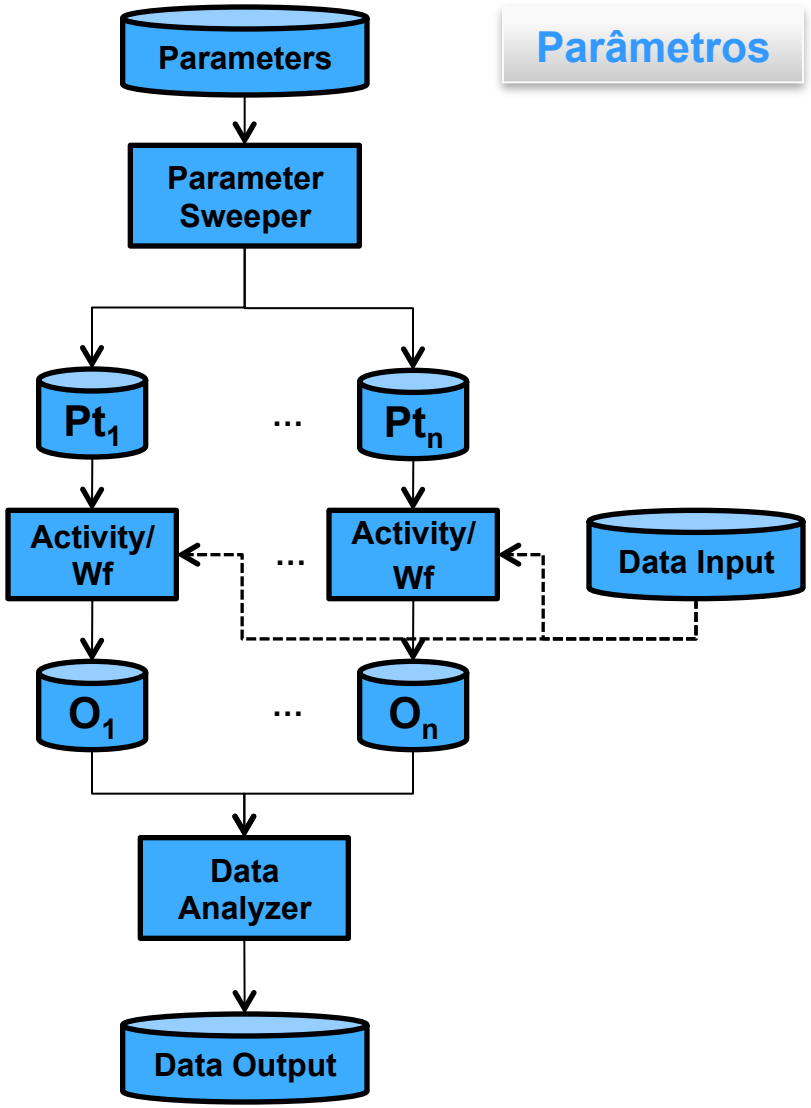
Workflow científico para previsão via redes neurais



Atividade distribuída no cluster

Tipos de paralelização





Dificuldades para paralelização

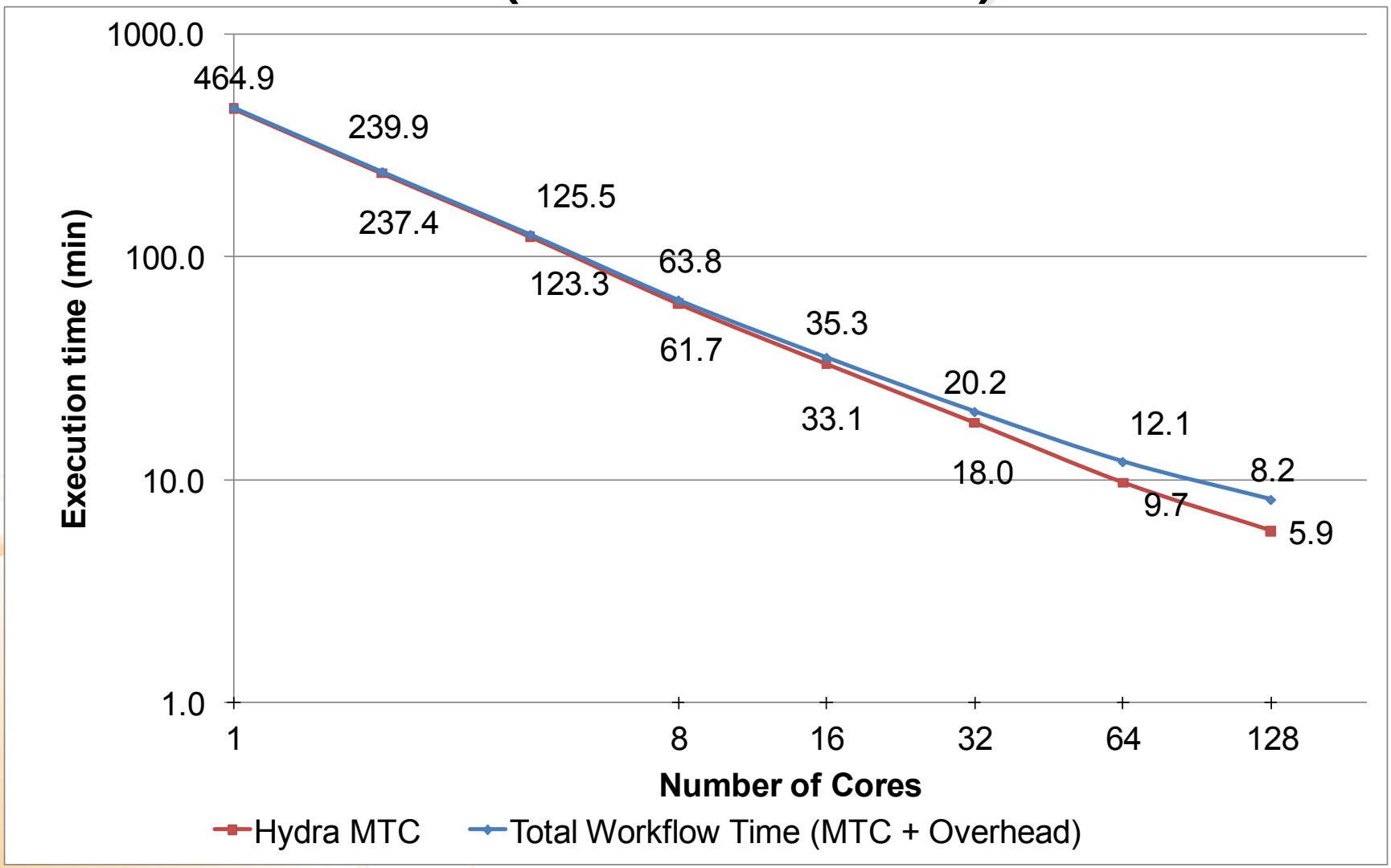
- Necessidade de aumento de desempenho de execução de atividades de um workflow
- Execução das atividades usando processamento de alto desempenho (*PAD*)
- SGWfC perde o controle da execução e desconhece estratégia de paralelismo
- Controle limitado da execução remota
- Proveniência limitada das atividades paralelas
- Exemplo: Workflow de treinamento das redes neurais

Uso do Hydra + VisTrails

- *Middleware para distribuição de atividades*
- Conjunto de componentes para um workflow científico
- Atividades podem ser paralelizadas
- O controle da execução fica com o SGWfC
- SGWfC conhece a estratégia de paralelismo
- Captura da proveniência das atividades paralelas
- Aumento de desempenho do workflow

OGASAWARA et al., 2009 - Exploring Many Task Computing in Scientific Workflows. In: ACM/IEEE SC09 - Workshop on Many-Task Computing on Grids and Supercomputers, 2009.

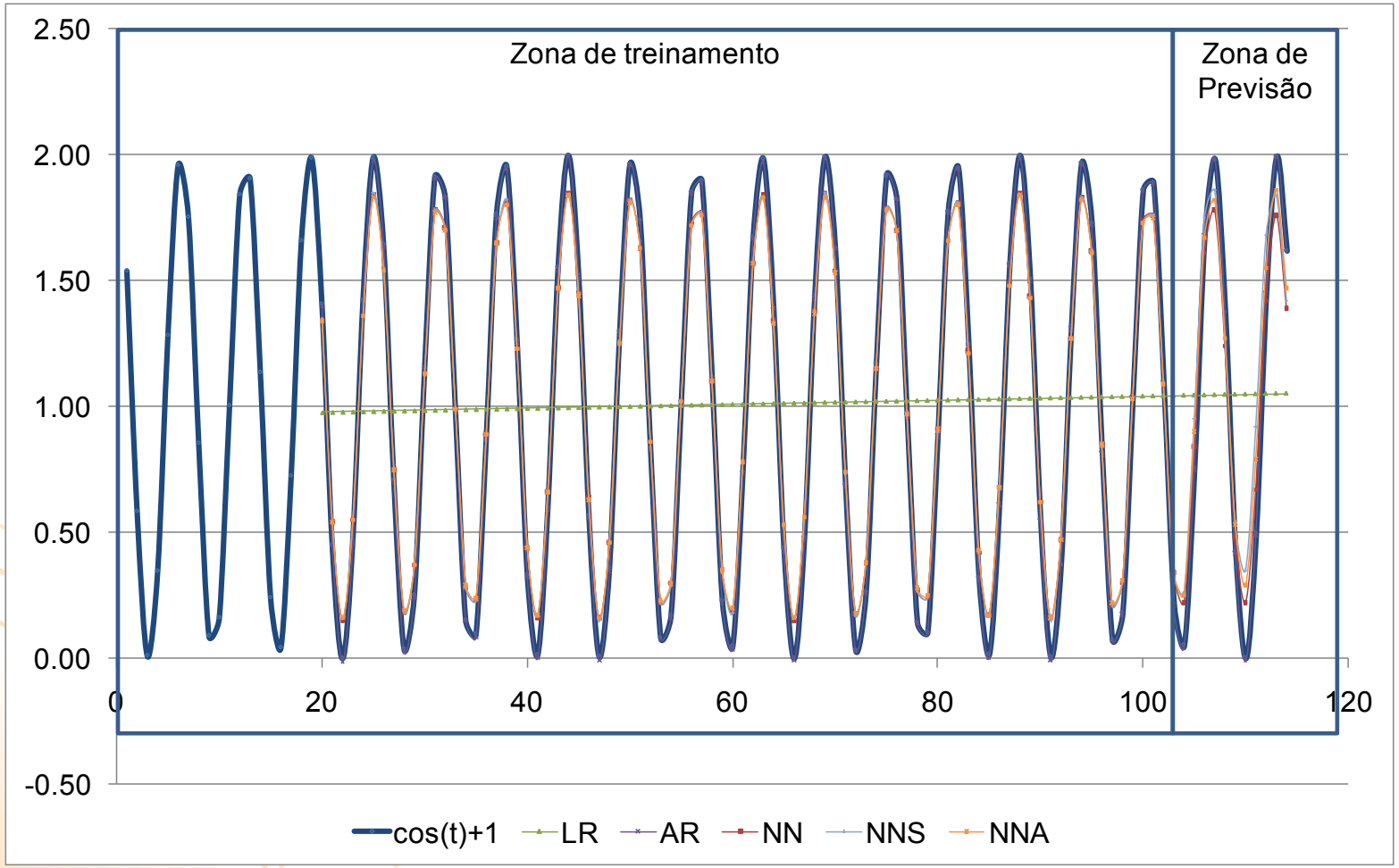
Resultados experimentais (velocidade)



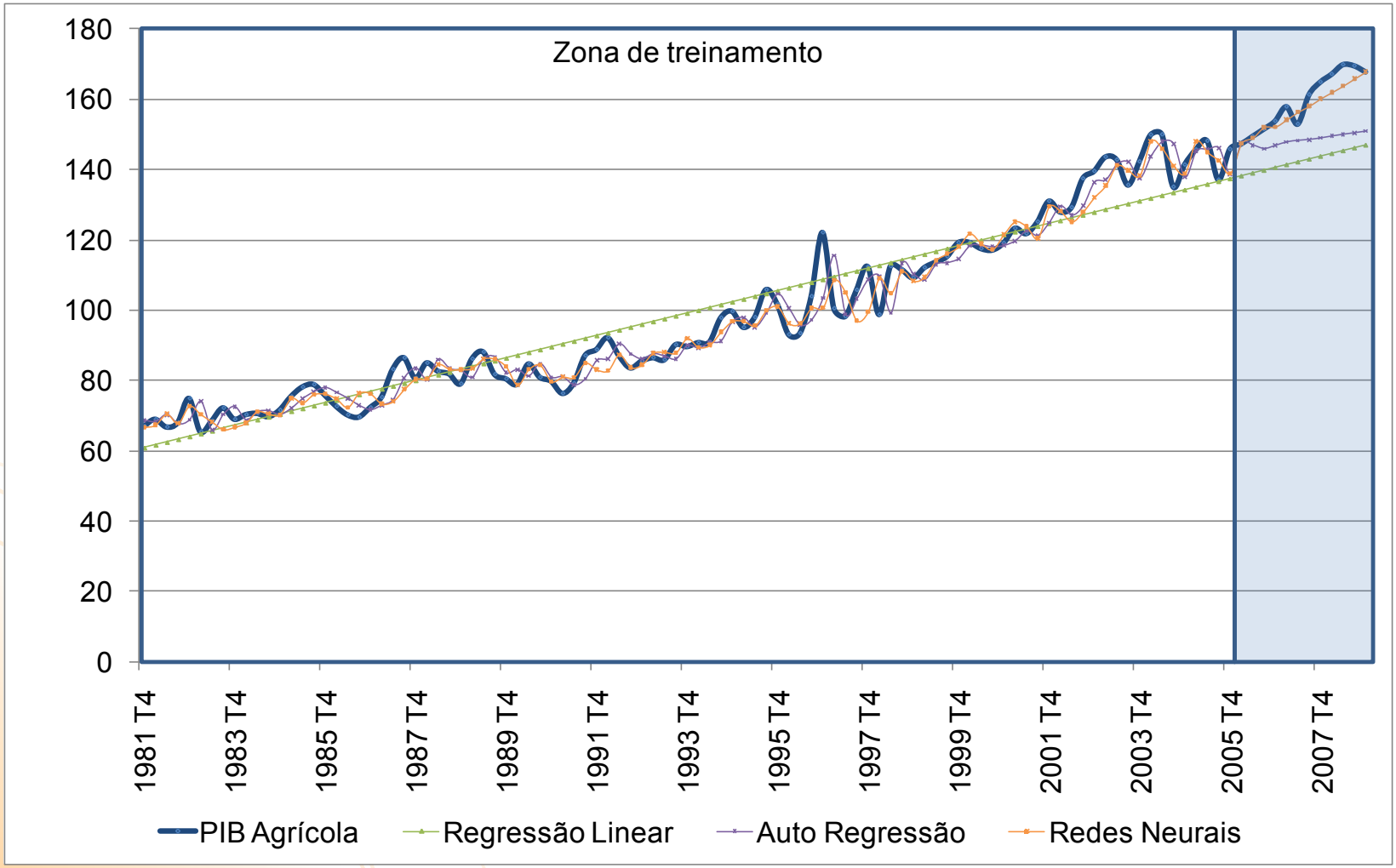
Resultados de proveniência

Number of activities (parameters explored)	512
Upload Transfer Time	1.2 min
Wait time for scheduler	0.3 min
Execution Time in HPC (16 nodes with 8 cores)	8.15 min
Download Transfer Time	0.8 min
Total Execution Time	**8.2 min**
Speedup	56.2
Number of errors	0

Resultados experimentais (previsão trivial)



Resultados experimentais (previsão do PIB agrícola)



Trabalhos relacionados

- *Weka – Knowledge Flow*
  - *Representa um fluxo de dados*
  - *Falta apoio a variações do experimento, a controles e a proveniência*
  - *Falta paralelismo*
- *Tamanduá*
  - *Projeto nacional para apoio a gestão governamental*
  - *Tem paralelismo*
  - *Falta apoio a variações do experimento, a controles e a proveniência*

Conclusões

- Viabilidade do uso de workflows científicos para obter paralelização
- Bom desempenho alcançado com pouco overhead do framework para paralelização
- Outras vantagens
  - Apoio na documentação / registro dos resultados obtidos
  - Facilidade na exploração
  - Facilidade na reprodução

Trabalhos Futuros

- Generalizar o mecanismo de agregação para ser utilizado em outros processos de mineração de dados
- Avaliar cenários de mineração de dados através de fragmentação de dados
- Avaliar o processo para os problemas de classificação

SBBD SBES 2009

V WAAMD

Paralelização de Tarefas de Mineração de Dados Utilizando Workflows Científicos

Obrigado!

Eduardo Ogasawara
ogasawara@cos.ufrj.br

Visite nosso sítio
http://gexp.nacad.ufrj.br

GExp

UFRJ COPPE

CNPq

PESC
Programa de Engenharia de Sistemas e Computação