Pesquisa em Algoritmos e Aplicações em Mineração de Dados no contexto do CEFET/RJ

Eduardo Ogasawara
eogasawara@cefet-rj.br

Centro Federal de Educação Tecnológica
Celso Suckow da Fonseca
CEFET/RJ

Agenda

- **Processo de mineração de dados**
- Visão geral do CEFET/RJ
- Pesquisa em Mineração de Dados

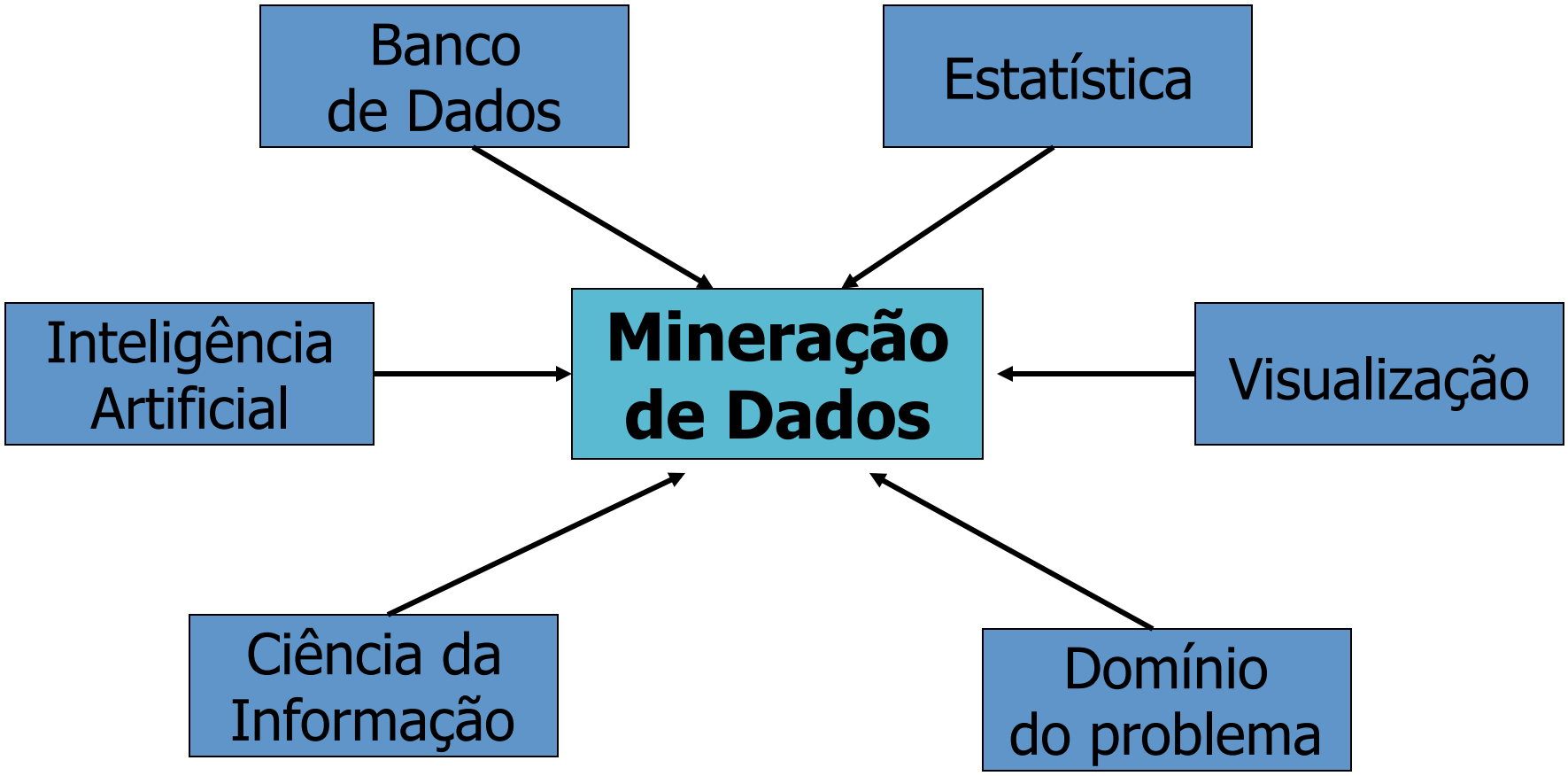
Visão Geral: Necessidade de extração de conhecimento

- **Explosão de dados (*data deluge*)**
  - **Coleta automatizada de dados por meio de ferramentas**
  - **Maturidade das tecnologias de banco de dados**
  - **Diferentes fontes: banco de dados e outros repositórios**
- **Estamos mergulhados em dados, mas famintos por conhecimento!**
- **Solução: *Data warehousing* e mineração de dados**
  - ***Data warehousing* e *on-line analytical processing* (OLAP)**
  - **Extração de conhecimento interessante (regras, padrões, restrições) a partir de "grandes" bases de dados**

Visão Geral: O que é mineração de dados?

- **Mineração de dados (MD)**
  - **Extração de informação ou padrões**
    - **não trivial**
    - **Implícita**
    - **previamente desconhecida**
    - **potencialmente útil**
- **Pode ser conhecida como:**
  - ***Knowledge discovery in databases (KDD)***
  - **Extração de conhecimento**
  - ***Business intelligence***
- **O que não é mineração de dados?**
  - **Processamento de consultas**
  - **Sistemas especialistas**
  - **Pacotes estatísticos**

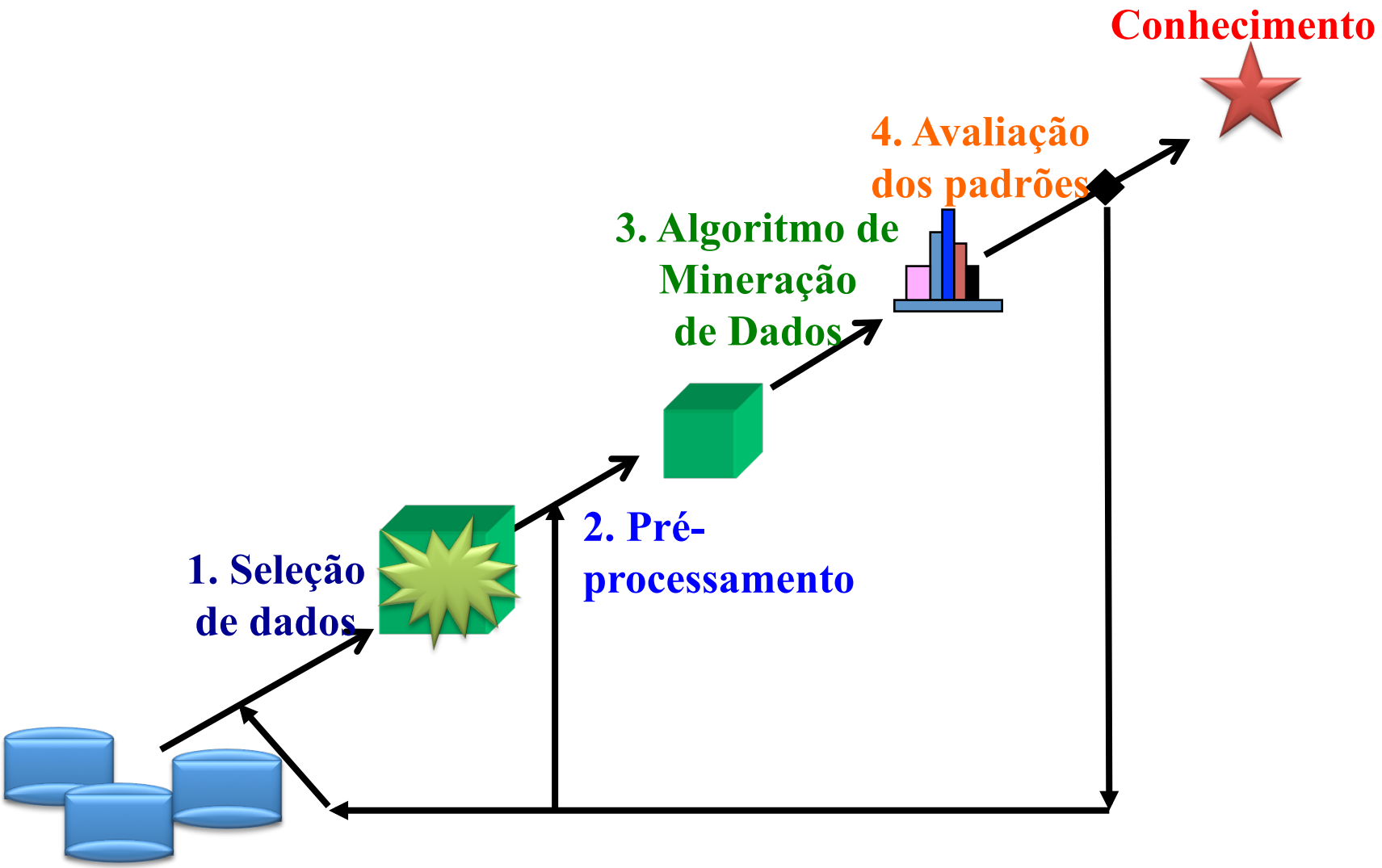
Visão Geral: Confluência de diversas disciplinas



Visão Geral: Responder as principais perguntas

- **Considere um exemplo de sistema de vendas pela internet**
- **Encontrei os dados?**
  - **Transações de cartões de crédito, cartões de fidelidade, central de atendimento ao consumidor, estilo de vida do cliente**
- **Consigo identificar grupos?**
  - **Identifiquei agrupamentos para "modelos" de clientes que compartilham algumas características: interesse, renda, hábitos de consumo**
- **Consigo classificar ou predizer?**
  - **Identifiquei padrões de consumo? Consigo estabelecer um padrão de consumo de solteiros e casados?**
- **Consigo associar?**
  - **Associação/correlação entre vendas de produtos**
  - **Predição baseada em associação de informação**

Processo de Mineração de Dados (PMD)



PMD: Passo 0: Aprendizado do domínio da aplicação

- **Tenha um conhecimento relevante e básico do domínio**
- **Identifique os objetivos da aplicação**
- **Avalie o problema na perspectiva do cliente**
- **Entenda os dados!**

PMD: Passo 1 – Identificação, Integração e Seleção

- **Identifique os tipos de dados**
  - **Espaciais, Temporais, Textuais, Multimídia, Grafos**
- **Identifique as fontes**
  - **Arquivos**
  - **Bases heterogêneas e legadas**
  - **Banco de dados relacional**
  - ***Data warehouses***
- **Integração de dados - *Data Warehouse***
  - **Integração dos dados provenientes de diferentes fontes**
    - **Os dados selecionados podem ir para um repositório**
  - **Importante, se for necessário realizar consultas OLAP**
- **Seleção**
  - **Criação de um data set alvo: seleção de dados**
  - **Seleção de um subconjunto de atributos**
  - **Seleção de um subconjunto de dados**

PMD: Passo 2 – Pré-processamento

- **Etapa extremamente importante**
  - **pode chegar a 60% do esforço!**
- **Redução de dados**
  - **Filtragem**
- **Redução de atributos**
  - **seleção de atributos úteis**
  - **redução da dimensionalidade**
- **Transformação**
  - **Combinação**
  - **Derivação**
  - **Agregação**

PMD: Passo 3 – Algoritmo de Mineração de Dados

- **Foco: Busca por padrões de interesse**
- **Escolha de abordagens de mineração de dados**
  - **agrupamento**
  - **classificação/previsão**
  - **associação**
  - **outras abordagens**
- **Escolha das técnicas**

PMD: Passo 4 – Avaliação de padrões

- **Medições de erro**
  - **Métricas de erro**
- **Avaliação de padrões e apresentação de conhecimento**
  - **Visualização**
  - **transformação**
  - **remoção de padrões redundantes**
- **Uso do conhecimento descoberto**
- **Será que todos os padrões "descobertos" são interessantes?**
  - **Um sistema de mineração de dados pode gerar milhares de padrões, mas nem todos são interessantes.**

Agenda

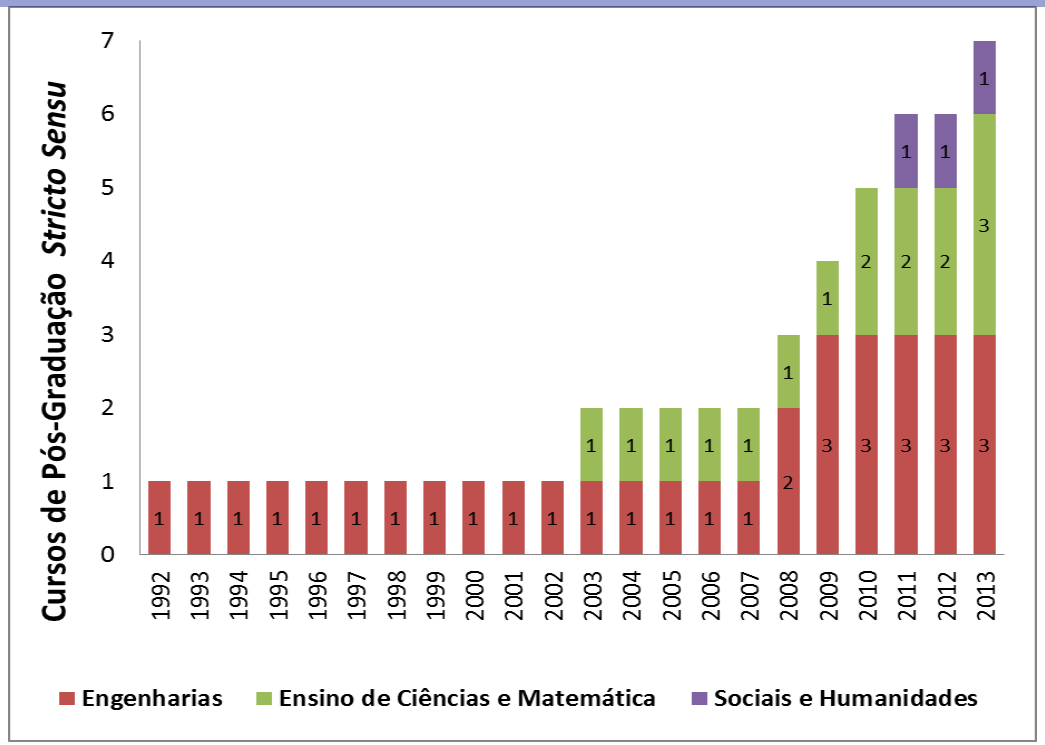
- Processo de mineração de dados
- **Visão geral do CEFET/RJ**
- Pesquisa em Mineração de Dados

Visão geral do CEFET/RJ

- ministrar educação profissional **técnica** de **nível médio** proporcionado habilitação aos setores da economia
- ministrar **ensino** superior de **graduação** e de **pós-graduação** *lato sensu* e *stricto sensu* em áreas científicas e tecnológicas
- realizar **pesquisa**, estimulando o desenvolvimento de soluções tecnológicas de forma criativa e estendendo seus benefícios à comunidade
- promover a **extensão** mediante integração com a comunidade, de modo que esta receba a transferência e aprimoramento dos benefícios e conquistas auferidos na atividade acadêmica e na **pesquisa aplicada**

Programas de Pós-Graduação Stricto Sensu

- **2** Doutorados
- **5** Mestrados **Acadêmicos**;
- **1** Mestrado **Profissional**;
- Mais de **400 dissertações** defendidas;
- **200 alunos** matriculados;
- **24 grupos de pesquisas ativos**;



Programas de Pós-Graduação *Stricto Sensu*

Tecnologia – PPTEC

Ensino de Ciências e Matemática – PPECM

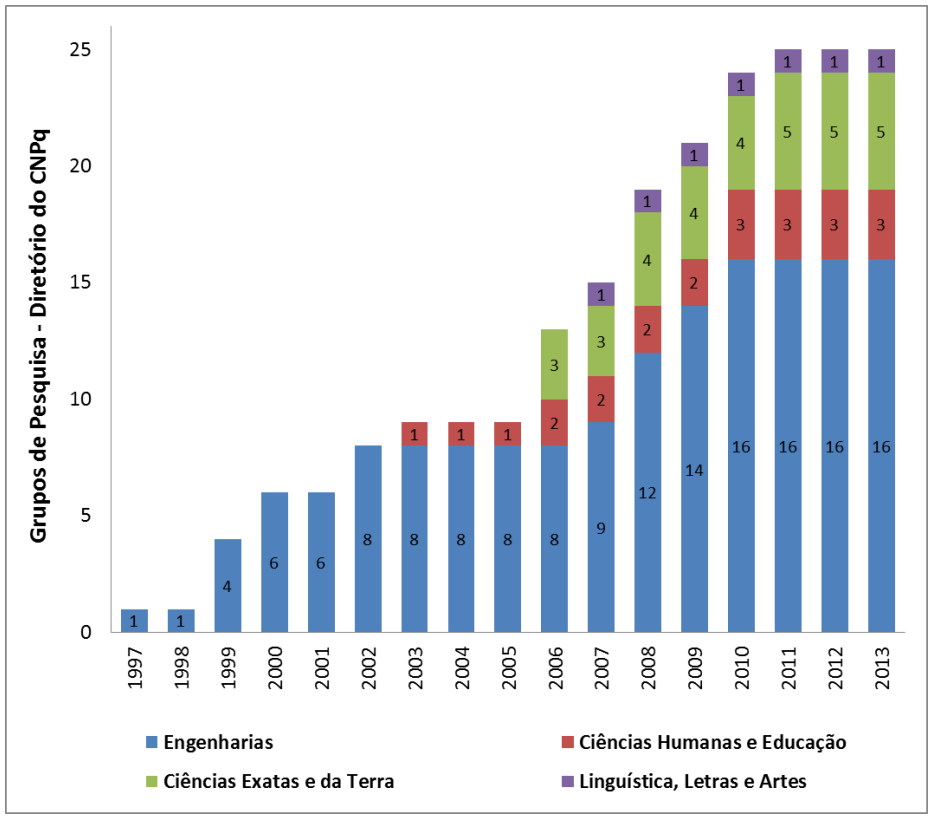
Engenharia Mecânica e Tecnologia dos Materiais – PPEMM

Engenharia Elétrica – PPEEL

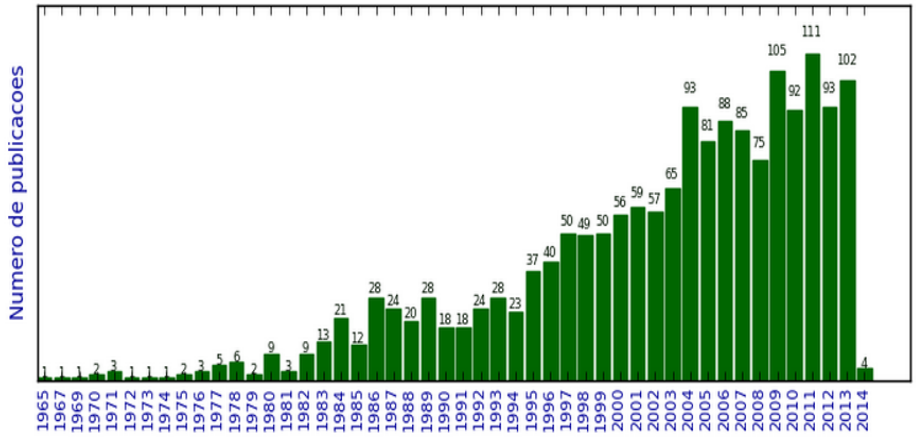
Ciência Tecnologia e Educação – PPCTE

Relações Etnicorraciais – PPRER

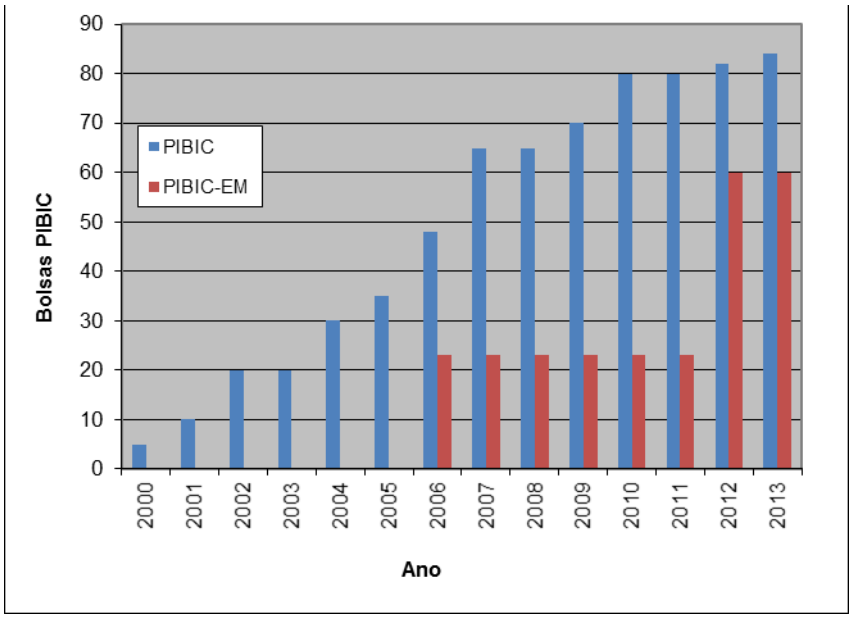
Pesquisa



Artigos completos publicados em periódicos



Número total de itens: 1699



Histórico do GPCA

- **Criado em 2011**
  http://dgp.cnpq.br/dgp/espelhogrupo/980693022019266 9
- **Ênfase em Computação Aplicada**
  - **Visão pragmática do mundo**
  - **Aplicabilidade em tudo o que é pesquisado**
- **Parcerias com demais centros de pesquisa**
  - **COPPE/UFRJ, UFF, LNCC, INMETRO, ON, UERJ, Unicamp, CETEM**
- **Parcerias com a indústria**
    - **Clavis, EUD, Fasolti**
- **Número de pesquisadores: 24**

Objetivos do GPCA

- Fortalecer a demanda pela criação do curso de Ciência da Computação ✔
- Servir como elo da Escola de Informática e Computação (EIC)
  - Fomentar pesquisa no Curso Técnico de Informática
  - Consolidar o curso de Ciência da Computação
  - Introduzir a pesquisa no Curso de Sistemas para Internet
- Posicionar a EIC em destaque na instituição
- Posicionar o CEFET/RJ em um contexto de relevância no cenário da computação nacional
- Servir como alicerce para os programas de pós-graduação do CEFET/RJ

Agenda

- Processo de mineração de dados
- Visão geral do CEFET/RJ
- **Pesquisa em Mineração de Dados**

Projeto de Pesquisa

- Objetivo
  - Desenvolvimento de algoritmos e aplicações de MD voltados a solução de problemas práticos
- Pesquisa
  - Básica
    - Produção de conhecimento ou ferramentas sem aplicação imediata a problemas reais
  - Aplicada
    - Uso do conhecimento ou ferramentas na resolução de problemas reais

Pesquisa básica em MD

- Algoritmos e Estrutura de Dados voltados a apoiar as etapas do processo de MD
  - Uniformização de dados e apoio na representação de experimentos por meio de uma abordagem algébrica
  - Trabalhos em etapas específicas do processo de mineração de dados
- Modelagem e execução desse processo por meio de workflows
  - Encadeamento de experimentos por meio de workflows de modo a execução em PAD

Aplicações de MD

- Aplicações em Ciência de Dados
  - Diferentes formas de gerir dados e algoritmos de mineração de dados em áreas que demandam grande exploração de dados (bioinformática, engenharia de produção, segurança da informação, redes sociais e astronomia)
- Aplicações Educacionais
  - Extração de conhecimento a partir de dados educacionais
  - Desenvolvimento de aplicativos de apoiam o ensino e produzem dados de uso
  - Desenvolvimento de repositórios para coleta de dados dos aplicativos
  - Processo de mineração de dados tradicional

Objetivos e Metas

- Identificação de problemas de MD formulados a partir de problemas práticos
- Aprimoramento do processo e das etapas de MD
- Aplicação de ferramentas de MD a problemas práticos

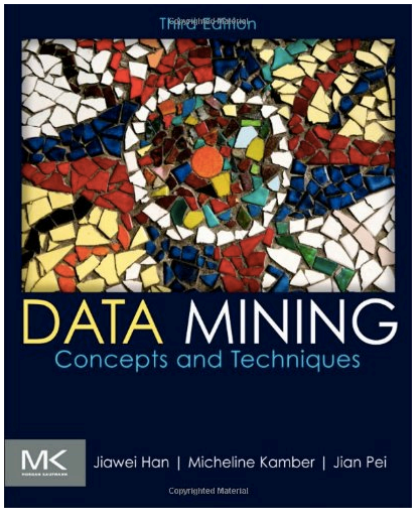
Trabalhos em andamento

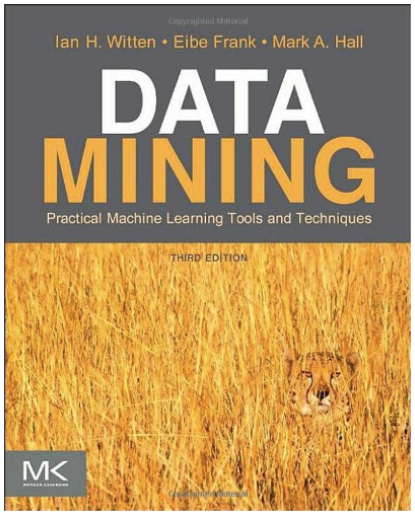
- **Forecasting of the Tropical Atlantic Sea Surface Temperatures through Time Series Frequency Change.**
  - **Trabalho que mostra que a troca de periodicidade das séries temporais na etapa de pré-processamento melhora o desempenho das previsões de longo prazo.**
- **Mercury: An Approach for Evaluating the Adoption of Enterprise Social Networks.**
  - **Trabalho que estabelece uma abordagem, constituída por métricas e avaliações estatísticas, para avaliar a adoção de redes sociais corporativas.**
- **Sagitarii: A Parallel, Distributed and Concurrent Workflow Engine for Data Mining.**
  - **Trabalho que estabelece uma abordagem algébrica para modelagem e execução de workflows de MD, de modo concorrente, em ambientes paralelos e distribuídos.**

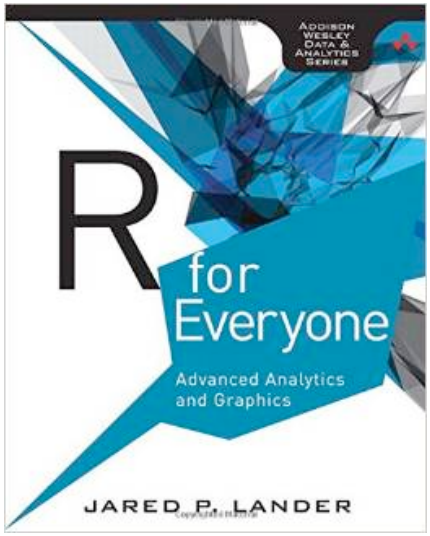
Trabalhos em andamento

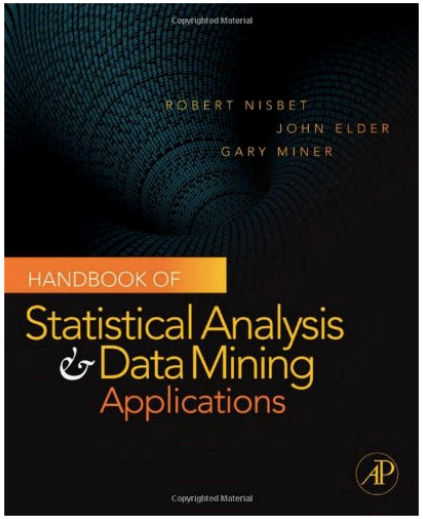
- **Top-k-MCluster: A top-k Multi-clustering Algorithm.**
  - **Trabalho que apresenta um algoritmo para produzir os *k* mais distintos agrupamentos formados a partir de um mesmo conjunto de dados.**
- **IDSFlow: A Set of Data Mining Workflows for Behavior-Based IDS.**
  - **Trabalho que apresenta um conjunto de workflows de MD para manutenção da eficiência de Sistemas de Detecção de Intrusão (do inglês, IDS) baseados em comportamento.**
- **Clustering of Flow Access from Web Servers.**
  - **Trabalho que apresenta um algoritmo para produzir agrupamentos de acessos temporais para identificação de padrões comportamentais de acesso a servidores web.**

Referências









Onde encontrar referências?

- **Específicas de mineração de dados:**
  - **Conferências: ACM SIGKDD, IEEE ICDM, SIAM-Data Mining, PAKDD, etc**
  - **Revistas: Data Mining and Knowledge Discovery, etc**
- **Na área de banco de dados:**
  - **Conferências: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, SBBD, etc**
  - **Revistas: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.**
- **Na área de IA:**
  - **Conferências: Machine learning, AAAI, IJCAI, etc.**
  - **Revistas: Machine Learning, Artificial Intelligence, etc.**