KDD E MINERAÇÃO DE DADOS:
POSICIONAMENTO E MOTIVAÇÃO

Prof. Paulo Mello

POSICIONAMENTO E MOTIVAÇÃO

- Coleta de dados em vários formatos, por meio de diversos recursos/aplicações em várias áreas:
  - Internet, dispositivos móveis, sensores, sistemas de automação, sistemas de informação, ...
  - Redes sociais, AVAs, redes de telecomunicações, operações com cartões de crédito, ...
  - Governo, (Bio)Ciências, Finanças, Seguros, Segurança, ...
  - IoT (Internet of Things – Internet das Coisas)
- Quanta informação é criada a cada ano?

POSICIONAMENTO E MOTIVAÇÃO

- Segundo a revista Science (2011): o mundo foi capaz de armazenar **295 exabytes** de informação no ano de **2007**.
  - 1 exabyte = 1012 megabytes
  - Cerca de 800 megabytes para cada ser humano.
  - Equivalente ao conteúdo textual de mais de 300 livros.
- **Atualmente** a NASA possui dados na ordem de **bilhões de gigabytes**.
- Estima-se que em **2020**, a humanidade disporá de **44 zettabytes** de dados.
  - 1 zettabyte = 44 trilhões de gigabytes (44×2^{70} bytes)
  - Taxa de crescimento de dados mundial em torno de 40% ao ano na próxima década.

Fontes:

www.sciencemag.org/content/early/2011/02/09/science.1200970.full.pdf

http://www.nasa.gov/open/plan/data-gov.html

www.emc.com/leadership/digital-universe/index.htm

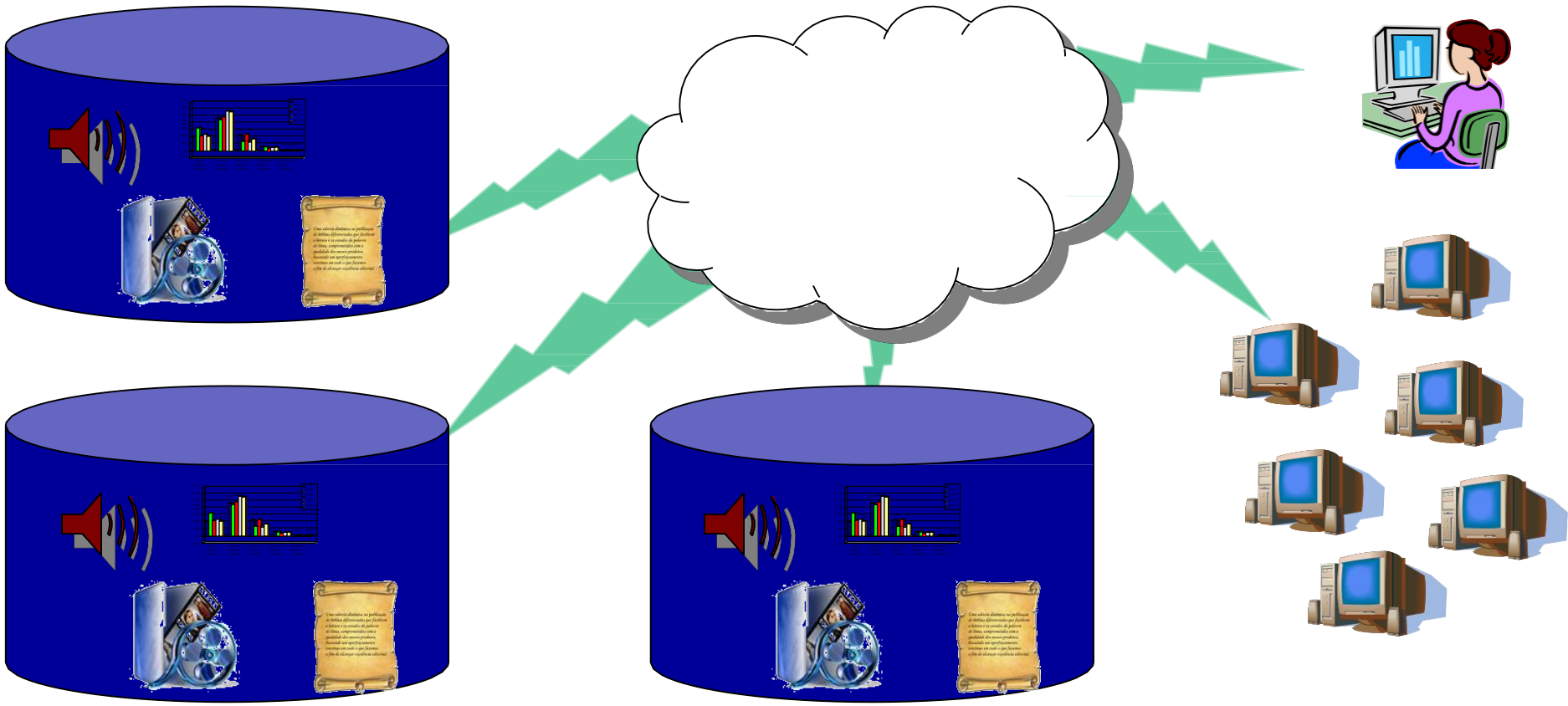
POSICIONAMENTO E MOTIVAÇÃO

- Nossa situação atual é a de **sobrecarga de informação**...

POSICIONAMENTO E MOTIVAÇÃO

Grandes Volumes de Dados Distribuídos

Vários formatos: texto, imagem, vídeos, sons, gráficos, etc...



POSICIONAMENTO E MOTIVAÇÃO

- Em vez de reduzir o problema, mecanismos de busca o amplificam, pois tornam novos documentos textuais rapidamente disponíveis.
- Muitos dados, pouca informação.
  - Google: 150M consultas/dia (2000/segundo)
  - Google: 4.2B documentos em seu índice
- Consequência: mais difícil extrair algo útil a partir dos dados (padrões, relacionamentos ou tendências subjacentes aos dados)
- A extração manual de informação é impossível.

Fonte: The Anatomy of a Large-Scale Hypertextual Web Search Engine
Sergey Brin and Lawrence Page, http://www-db.stanford.edu/~backrub/google.html

POSICIONAMENTO E MOTIVAÇÃO

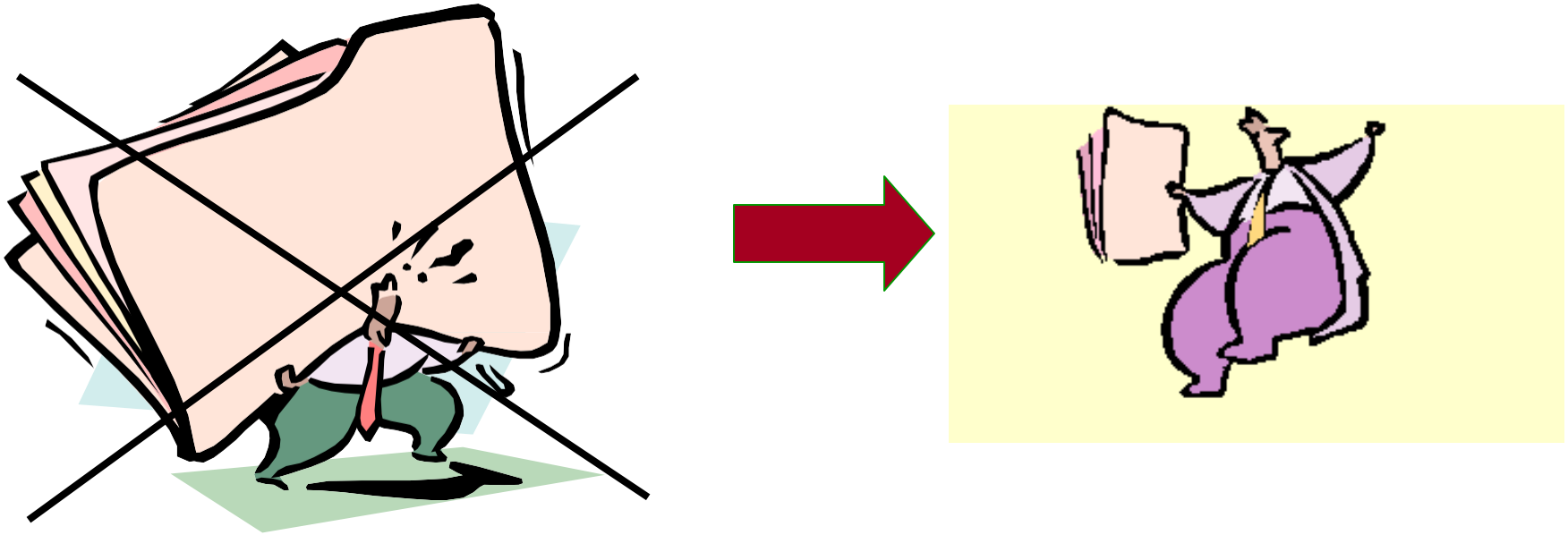
Exemplos de Instituições com BDs Massivos:

- **FedEx**
- **UPS**
- **Wal-Mart**
- **NASA**
- **Projeto Genoma**
- **Caixa Econômica**
- **Banco do Brasil**
- **Dentre muitos outros …**

POSICIONAMENTO E MOTIVAÇÃO

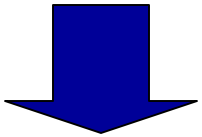
Necessidade:

Ferramentas **inteligentes** que auxiliem na **análise de dados** e na **busca por conhecimentos** em **GRANDES** conjuntos de dados (nos mais **diversos formatos**).

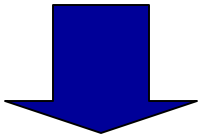


POSICIONAMENTO E MOTIVAÇÃO

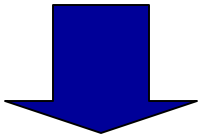
Avanços em TI



Crescimento Exponencial de BDs



Necessidade de Ferramentas para Análise Grandes BDs

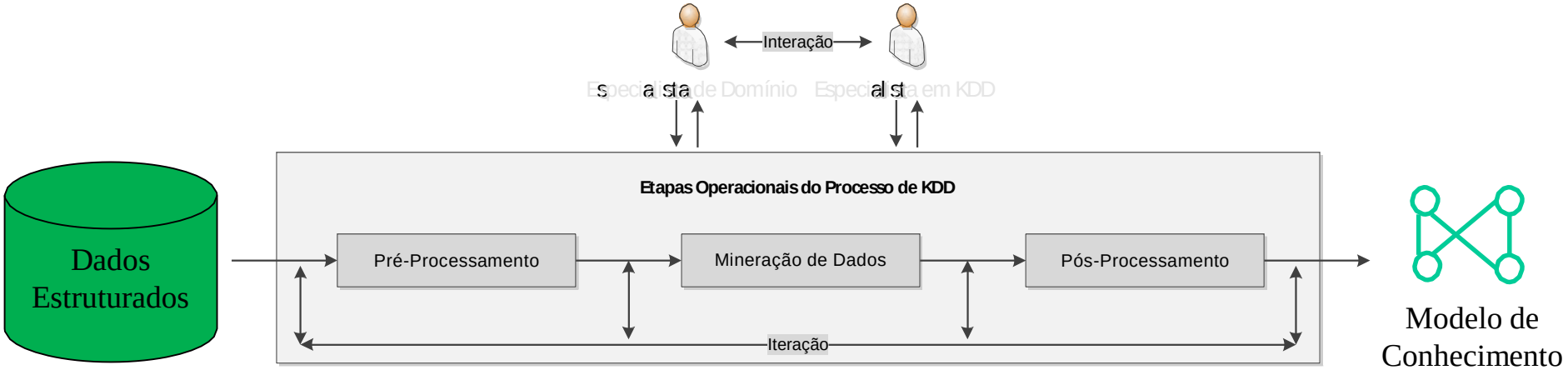


Área da Descoberta do Conhecimento em Bases de Dados (KDD)

POSICIONAMENTO E MOTIVAÇÃO

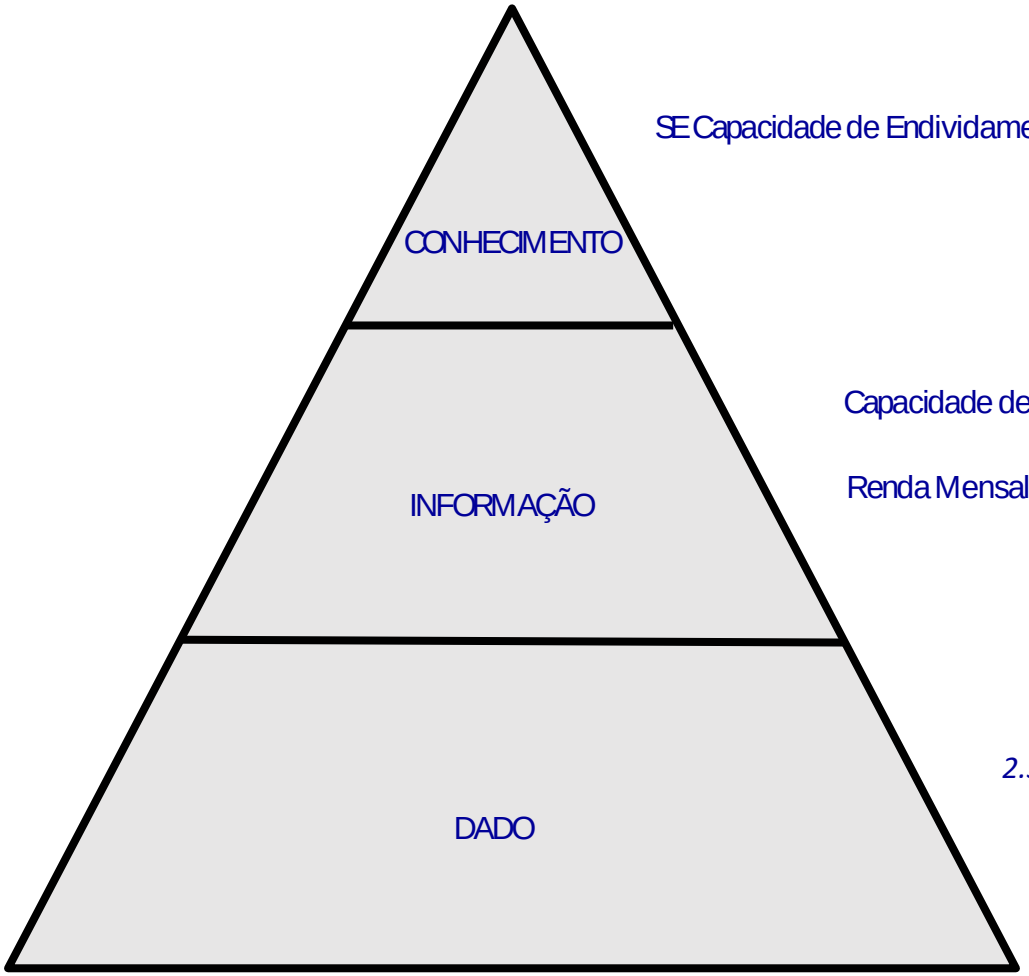
Descoberta de Conhecimento em Bases de Dados – KDD

"É um **processo**, de várias etapas**,** não trivial, **interativo e iterativo**, para **identificação** de **padrões compreensíveis**, **válidos**, **novos** e potencialmente **úteis** a partir de grandes conjuntos de dados." **[Fayyad et al., 1996]**



POSICIONAMENTO E MOTIVAÇÃO

Hierarquia Dado - Informação - Conhecimento:



SE Capacidade de Endividamento Mensal > 0.6 ENTÃO Crédito = Sim

Capacidade de Endividamento Mensal = 1 – Despesa Mensal / Renda Mensal

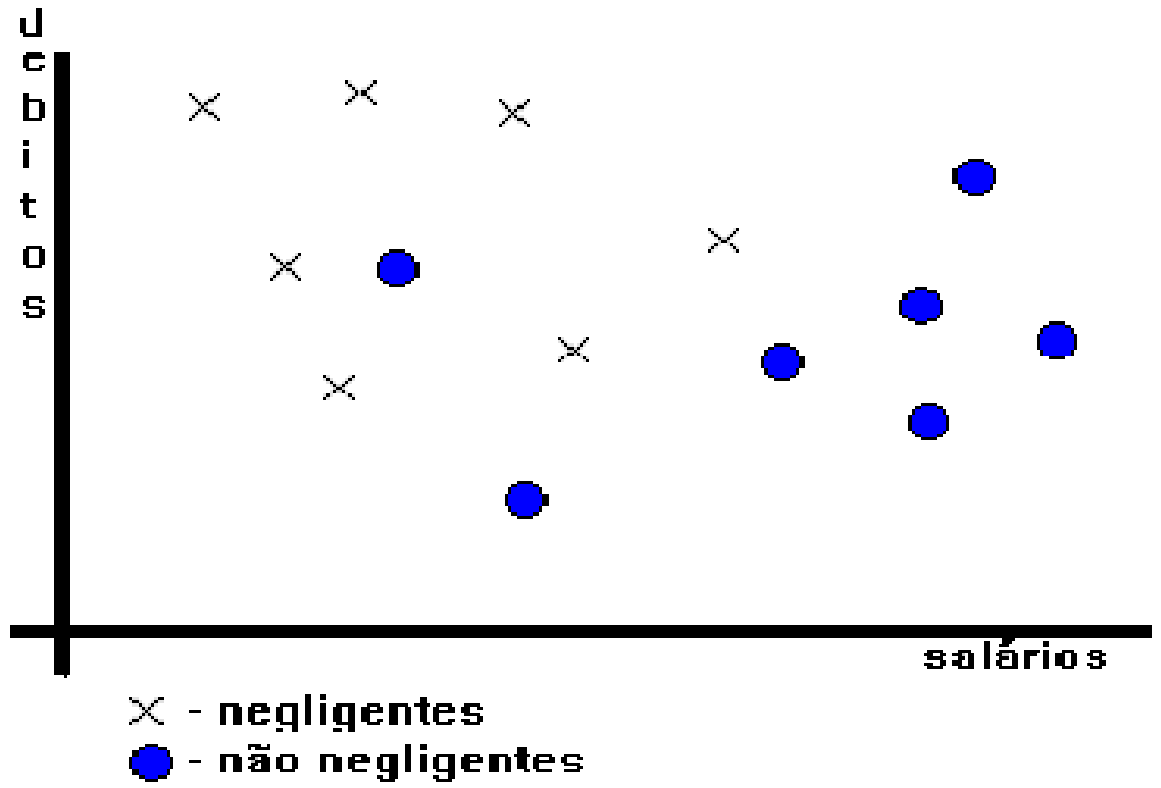
Renda Mensal, Despesa Mensal

2.345,20; 463,00; 10.048,21; 294,12

POSICIONAMENTO E MOTIVAÇÃO

Exemplo de aplicação de KDD na área de concessão de crédito:

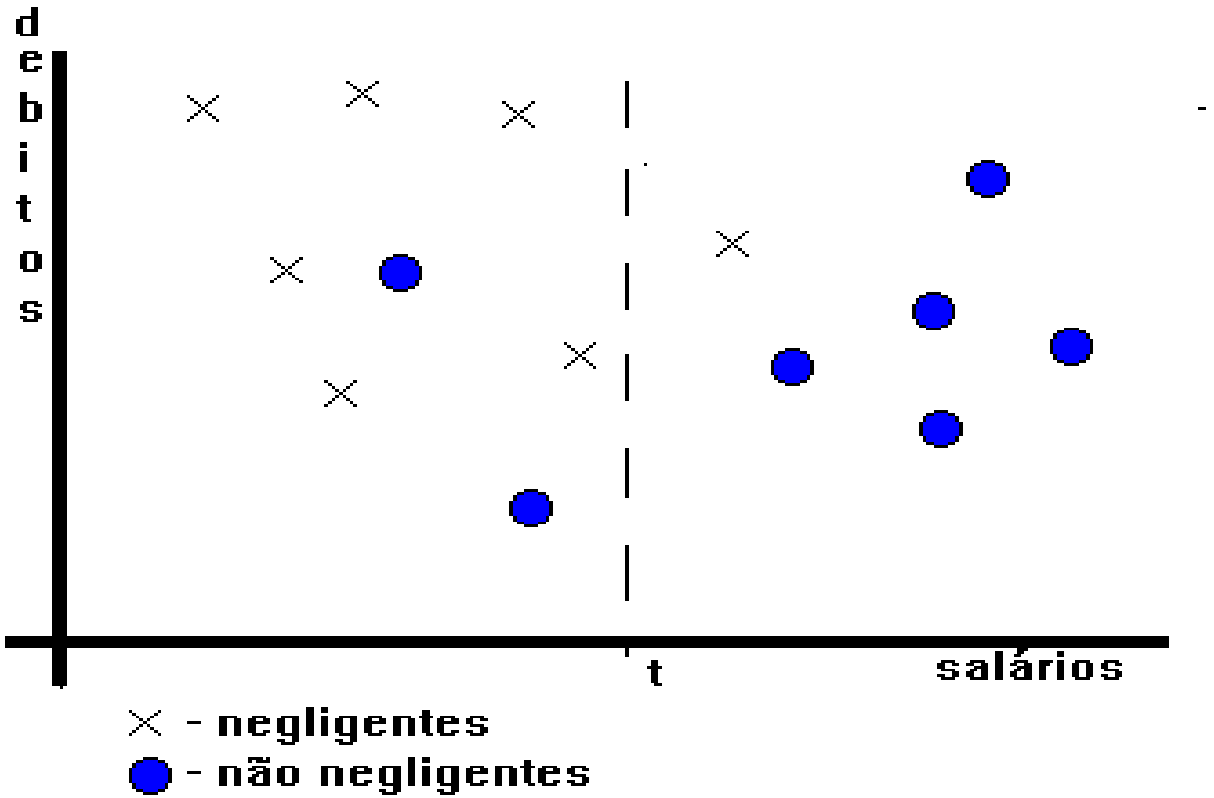
Conjunto de dados (Fatos)



POSICIONAMENTO E MOTIVAÇÃO

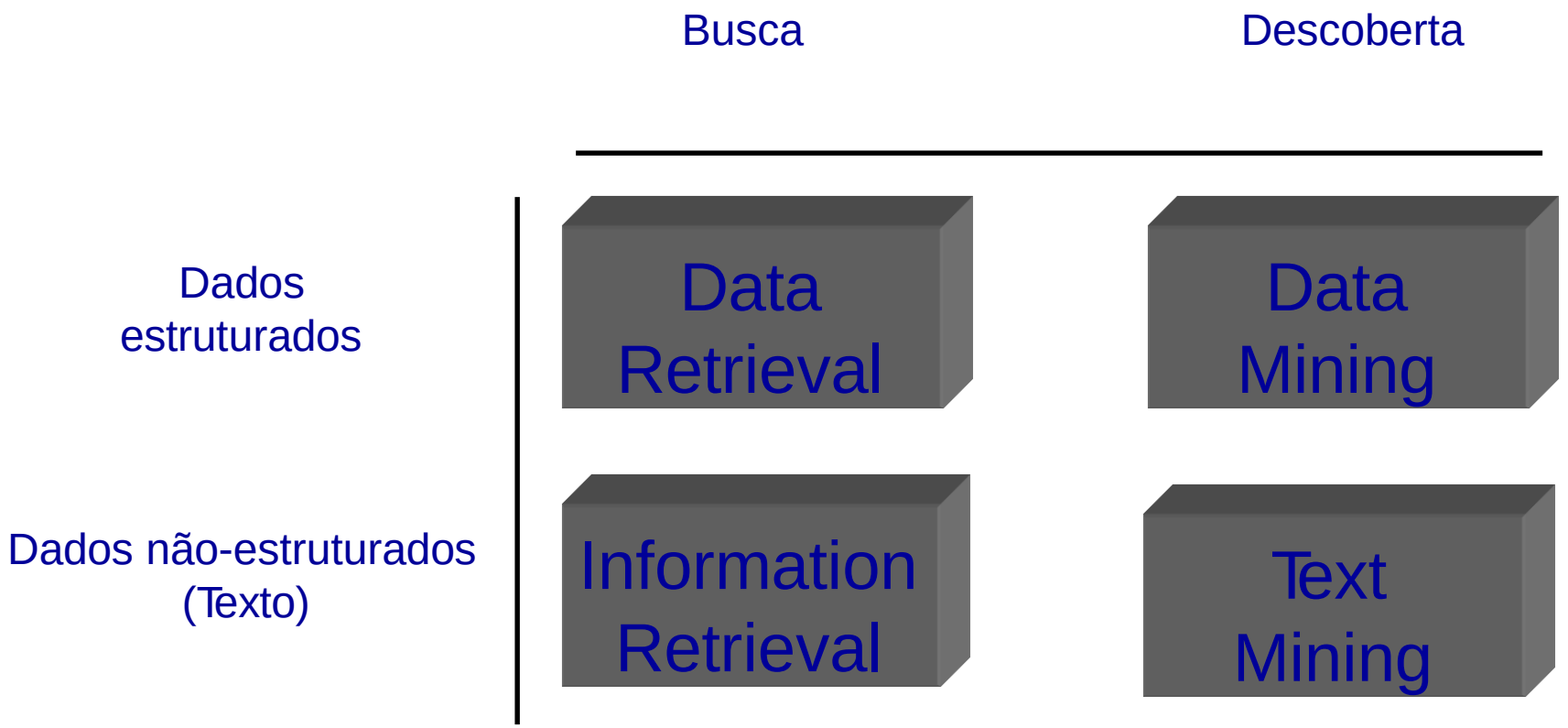
Exemplo de aplicação de KDD na área de concessão de crédito:

Padrão: Se renda > R$ t Então Crédito = SIM (Cto)



POSICIONAMENTO E MOTIVAÇÃO

"BUSCA" VS "DESCOBERTA"



POSICIONAMENTO E MOTIVAÇÃO

- Na verdade, há vários tipos de “mining”, dependendo da natureza dos dados:
  - Data Mining
  - Web Mining
    - Conteúdo
    - Estrutura
    - Log dos servidores
  - Multimídia Mining (Som, Imagem, …)
  - Text Mining

- Terminologia acima não é um consenso.

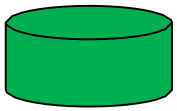
POSICIONAMENTO E MOTIVAÇÃO

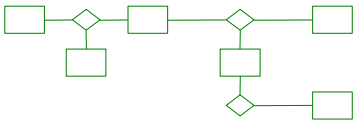
DESCOBERTA DE CONHECIMENTO - UMA TAXONOMIA

Descoberta de Conhecimento

Dados Estruturados

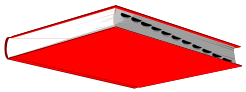
Mineração de Dados Clássica





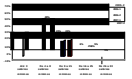
Dados Textuais

Mineração de Dados Textuais



Dados Multimídia

Mineração de Dados Multimídia

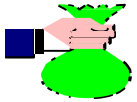


POSICIONAMENTO E MOTIVAÇÃO

Áreas de Origem

POSICIONAMENTO E MOTIVAÇÃO

EXEMPLOS DE ÁREAS DE APLICAÇÃO:

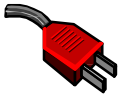


Comércio

Finanças

Medicina

Educação



Energia

Telecomunicações

Meio-Ambiente

Indústria

Etc...

POSICIONAMENTO E MOTIVAÇÃO

- **Comércio / Marketing**

  Perfil do Consumidor (Marketing Direto), Promoção de Produtos, Segmentação de Mercado, etc;...

- **Finanças**

  Análise de Investimentos, Análise de Crédito, Detecção de Fraudes em compras de Cartão de Crédito, etc;...

- **Medicina**

  Diagnóstico e Prevenção de Doenças, Detecção de Fraudes em Planos de Saúde, etc;...

POSICIONAMENTO E MOTIVAÇÃO

- **Educação**

  Análise de Matrículas e Demandas por Escolas, Evasão Escolar, Um Computador por Aluno;...

- **Energia**

  Previsão de Demanda, Distribuição de Recursos;...

- **Telecomunicações**

  Detecção de falhas, Dimensionamento de Sistemas de Comunicação, Detecção de Fraudes;...

POSICIONAMENTO E MOTIVAÇÃO

- **Meio Ambiente**

  Monitoramento ambiental, Prevenção de desequilíbrios ecológicos;...

- **Indústria**

  Previsão de demanda, Planejamento da produção e distribuição;...

- **Área Social**

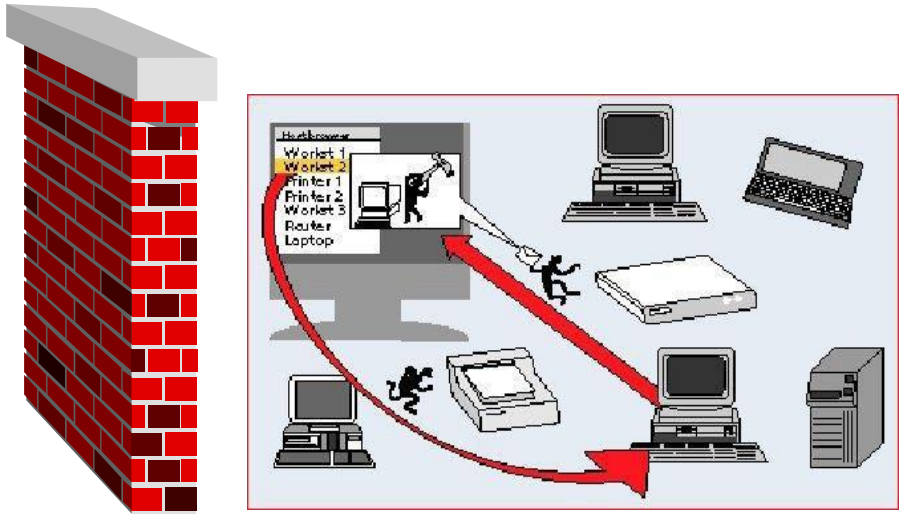
  Caracterização de Perfil para Reintegração Social;...

POSICIONAMENTO E MOTIVAÇÃO

Exemplos na área da Segurança

Como saber se uma mensagem é lixo ou de fato interessa?

Como saber se um dado comportamento de usuário é suspeito e com lidar com isto?



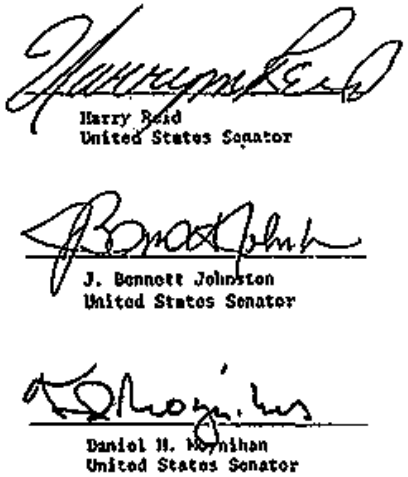
Detecção de intrusão e filtragem de spam

POSICIONAMENTO E MOTIVAÇÃO

Exemplos de aplicação de Mineração de Dados:

Classificação de imagens baseada em conteúdo

Identificação por impressões digitais

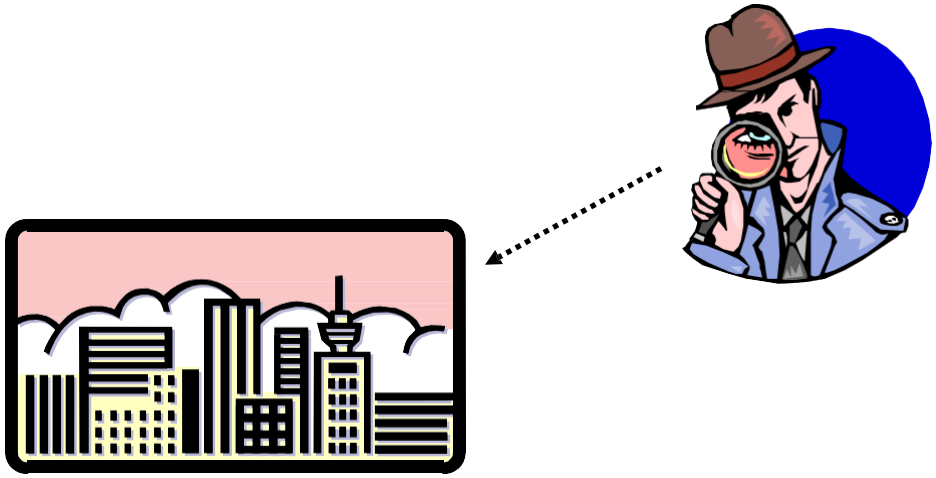


Reconhecimento automático de assinaturas

POSICIONAMENTO E MOTIVAÇÃO

Exemplos de aplicação de Mineração de Dados:

Classificação de imagens baseada em conteúdo



- Autêntico

ou

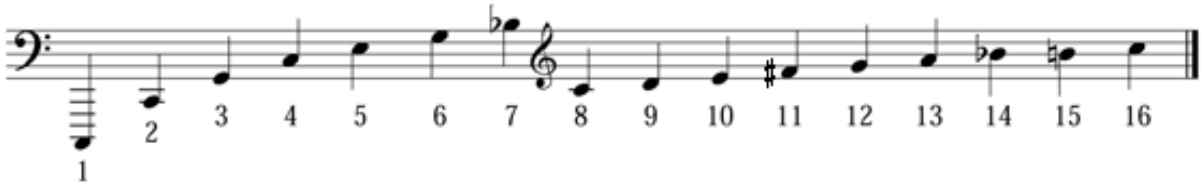
- Fraude

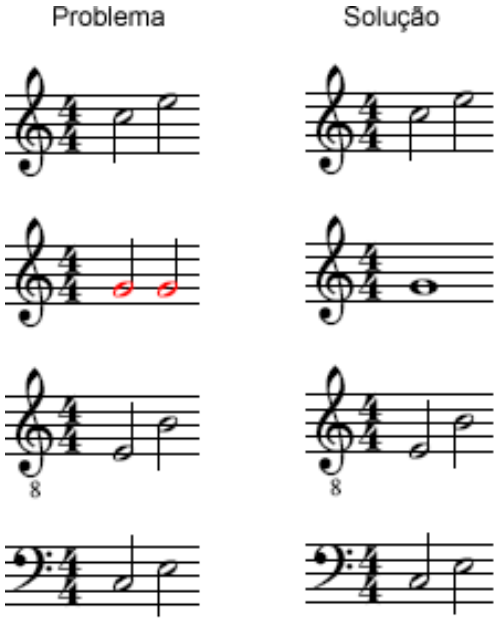
Projeto PORTINARI

POSICIONAMENTO E MOTIVAÇÃO

Exemplos de aplicação de Mineração de Dados:

Extração e correção de padrões em músicas





POSICIONAMENTO E MOTIVAÇÃO

Exemplos de aplicação de Mineração de Dados:

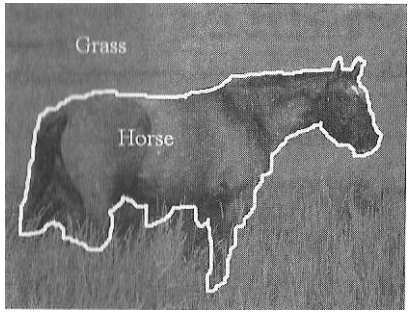
Reconhecimento e classificação de sons

Reconhecimento de Voz e de Locutores

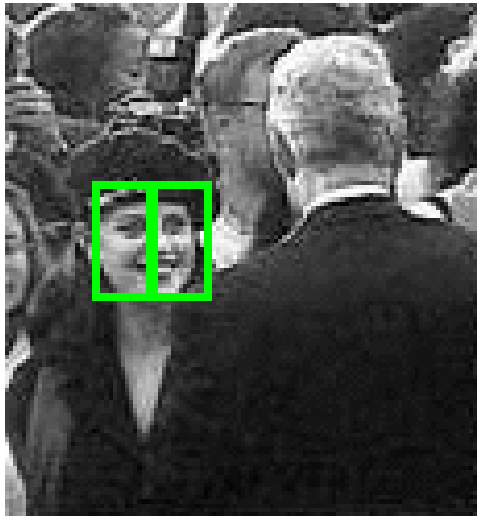
POSICIONAMENTO E MOTIVAÇÃO

Exemplos de aplicação de Mineração de Dados:

Reconhecimento e busca de objetos em imagens ou vídeos



Identificação de Elementos



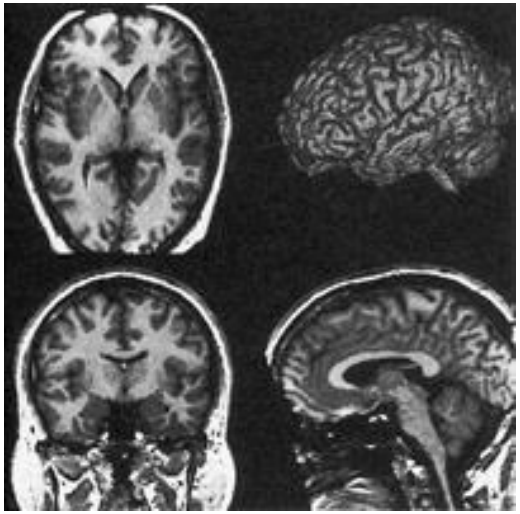
Reconhecimento de face

POSICIONAMENTO E MOTIVAÇÃO

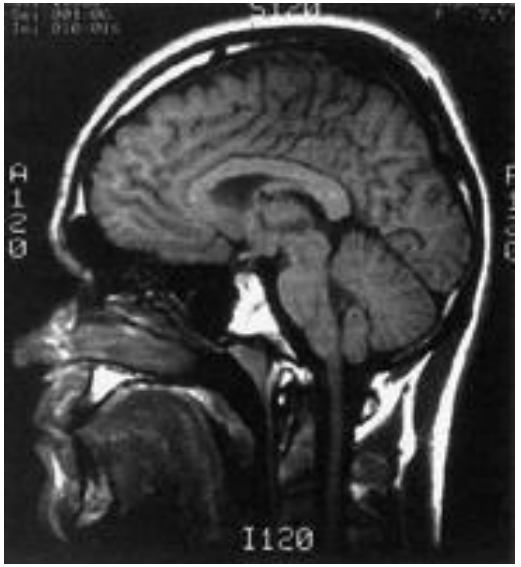
Exemplos de aplicação de Mineração de Dados:

Reconhecimento e busca de objetos em imagens ou vídeos

Diagnóstico a partir de radiografia



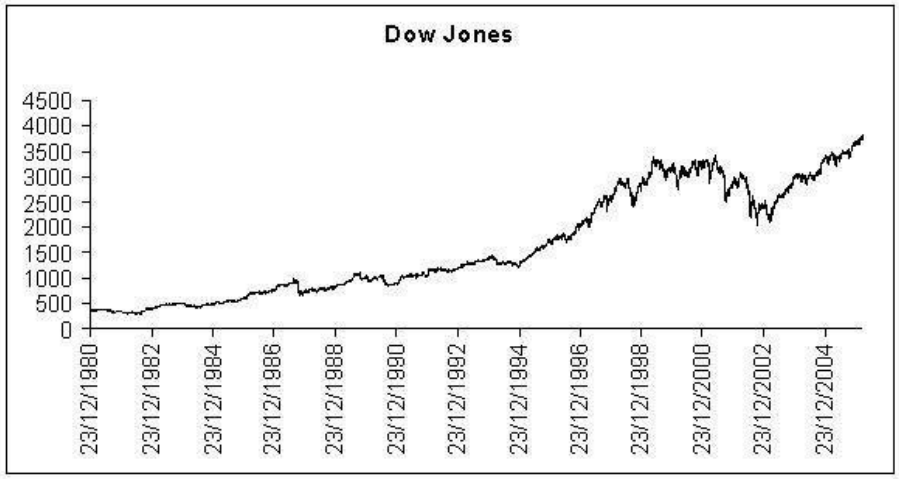
Diagnóstico a partir de tomografia computadorizada



Diagnóstico a partir de ressonância magnética

POSICIONAMENTO E MOTIVAÇÃO

Exemplos na área Financeira



Previsão da cotação de ações na bolsa de valores

POSICIONAMENTO E MOTIVAÇÃO

Exemplos na área de Energia (Petróleo)

Fotos Originais:

Fotos com tratamento da Luminosidade:

Fotos com tratamento do Filtro Gamma:

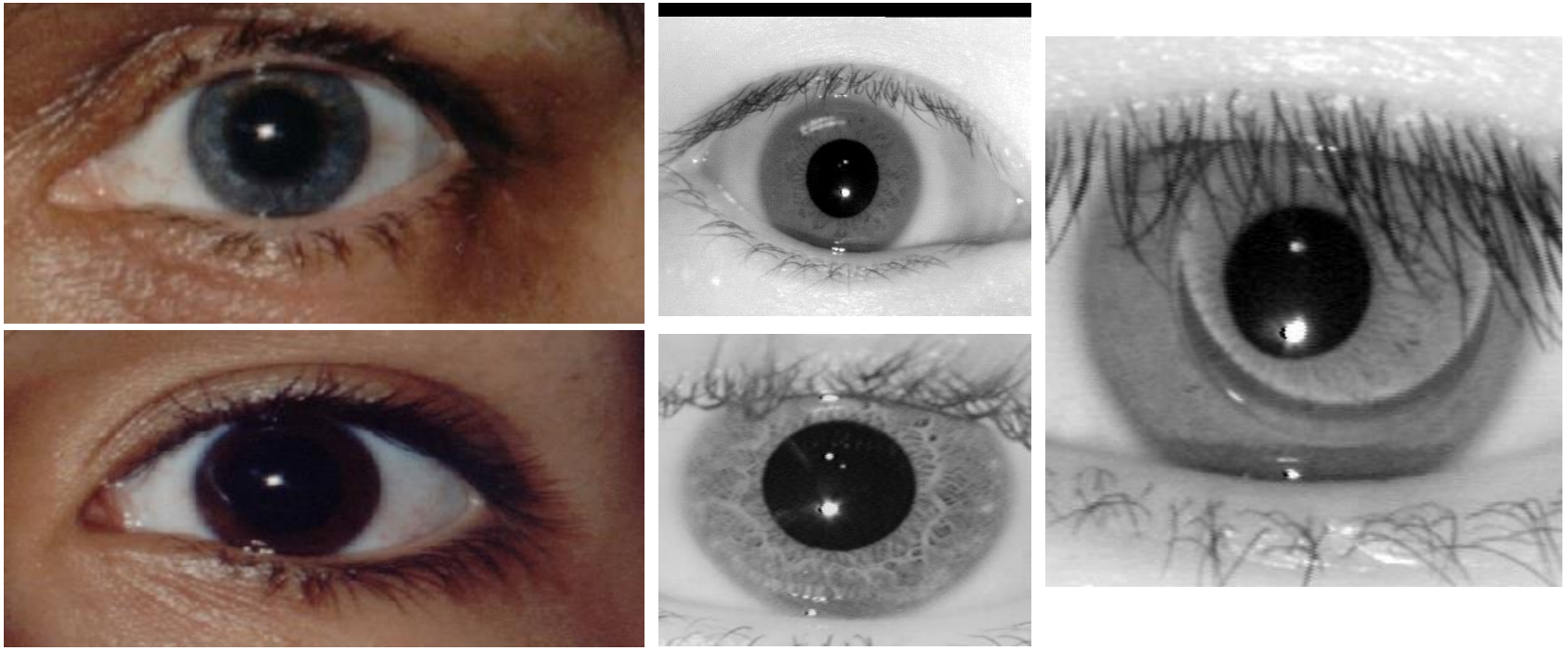
Fotos segmentadas e pós-processadas

Identificação de locais para perfuração de poços de petróleo

POSICIONAMENTO E MOTIVAÇÃO

Exemplos de aplicação de Mineração de Dados:

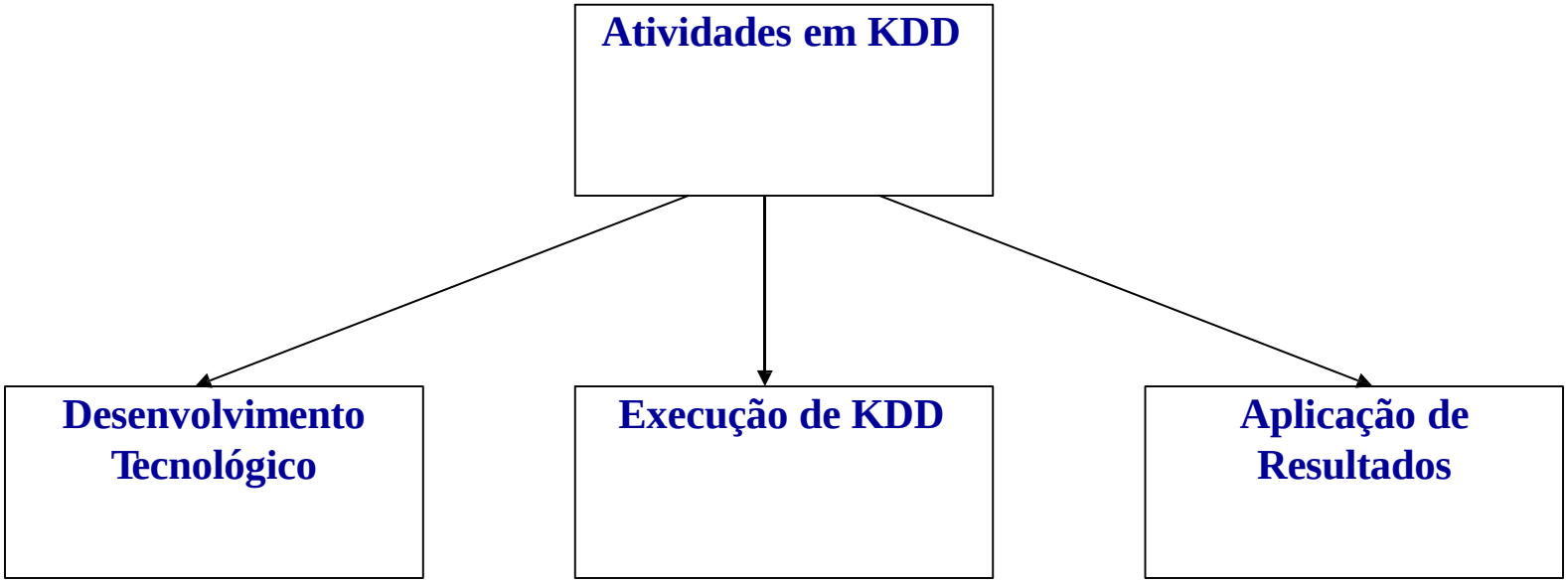
Reconhecimento de imagens baseada em conteúdo



Reconhecimento de usuários pela íris

POSICIONAMENTO E MOTIVAÇÃO

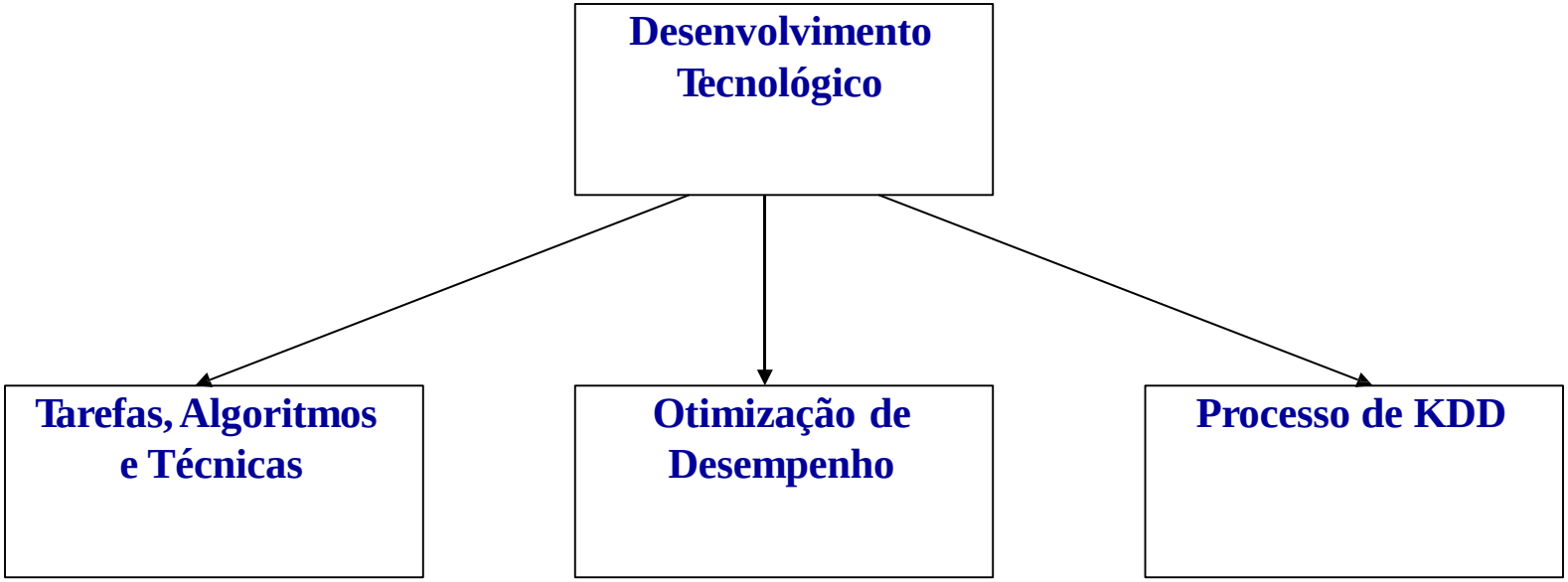
Atividades em KDD - uma Taxonomia



[Goldschmidt et al., 2002a]

POSICIONAMENTO E MOTIVAÇÃO

Atividades em KDD - uma Taxonomia



[Goldschmidt et al., 2002a]

POSICIONAMENTO E MOTIVAÇÃO

Tópicos Relacionados:

- Mineração de Textos
- Mineração de Dados Multimídia
- Mineração de Grafos
- Big Data
- Mineração de Dados Paralela e Distribuída

POSICIONAMENTO E MOTIVAÇÃO

Tópicos Relacionados:

- *Opinion Mining*
- *Educational Data Mining*
- *Social Data Mining*
- *Web Mining*
- *Etc…*