Mineração de Dados com Apache Mahout

P&D – 26/06/2013

by

Fabíola Souza Fernandes Pereira

Walmart

The Financial Times of London (1996)

Walmart

Fraldas e cerveja?

The Financial Times of London (1996)

TECH | 2/16/2012 @ 11:02AM | 564,678 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

The NY Times (2012)

Grávida?

TECH | 2/16/2012 @ 11:02AM | 564,678 views

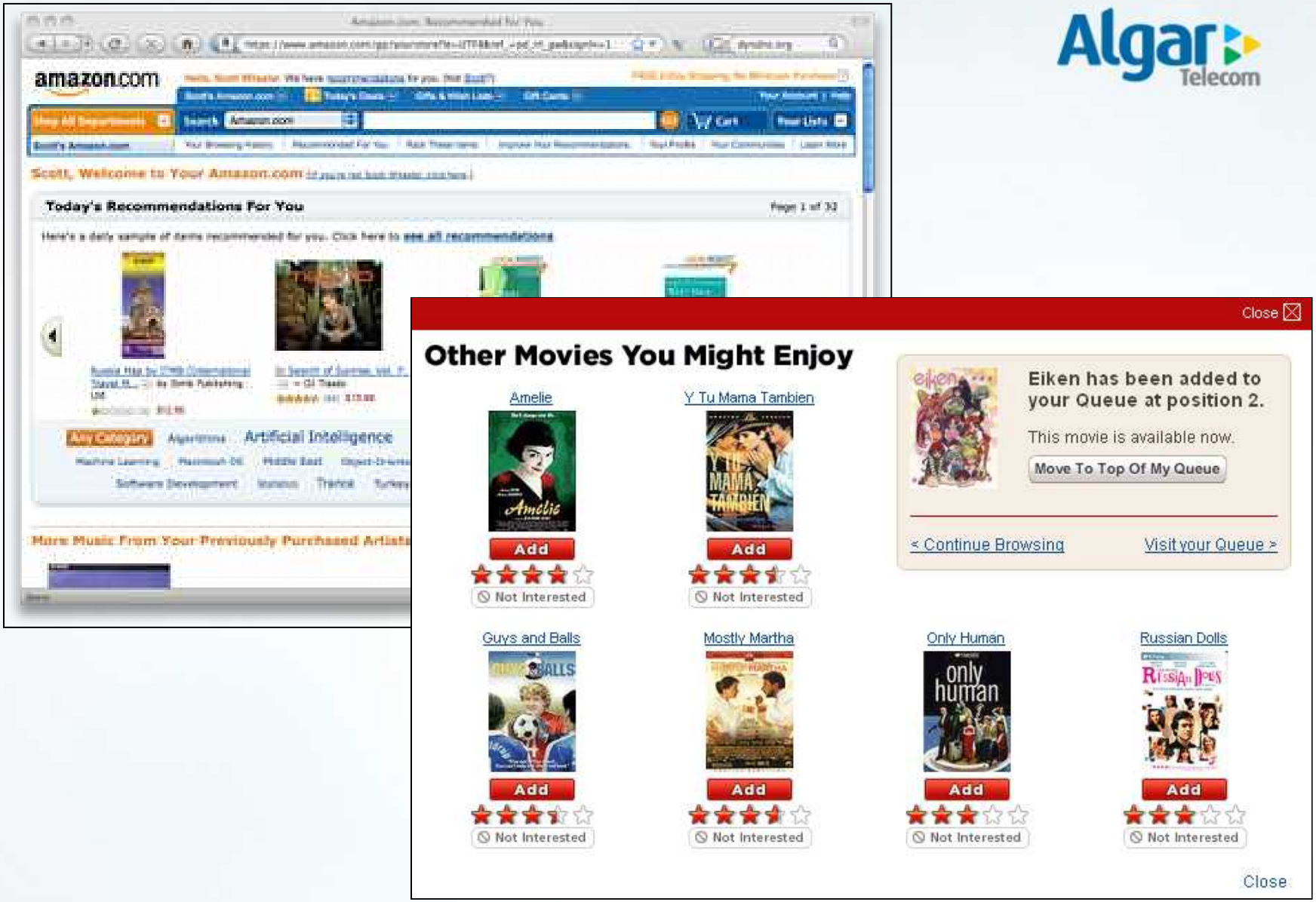
How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

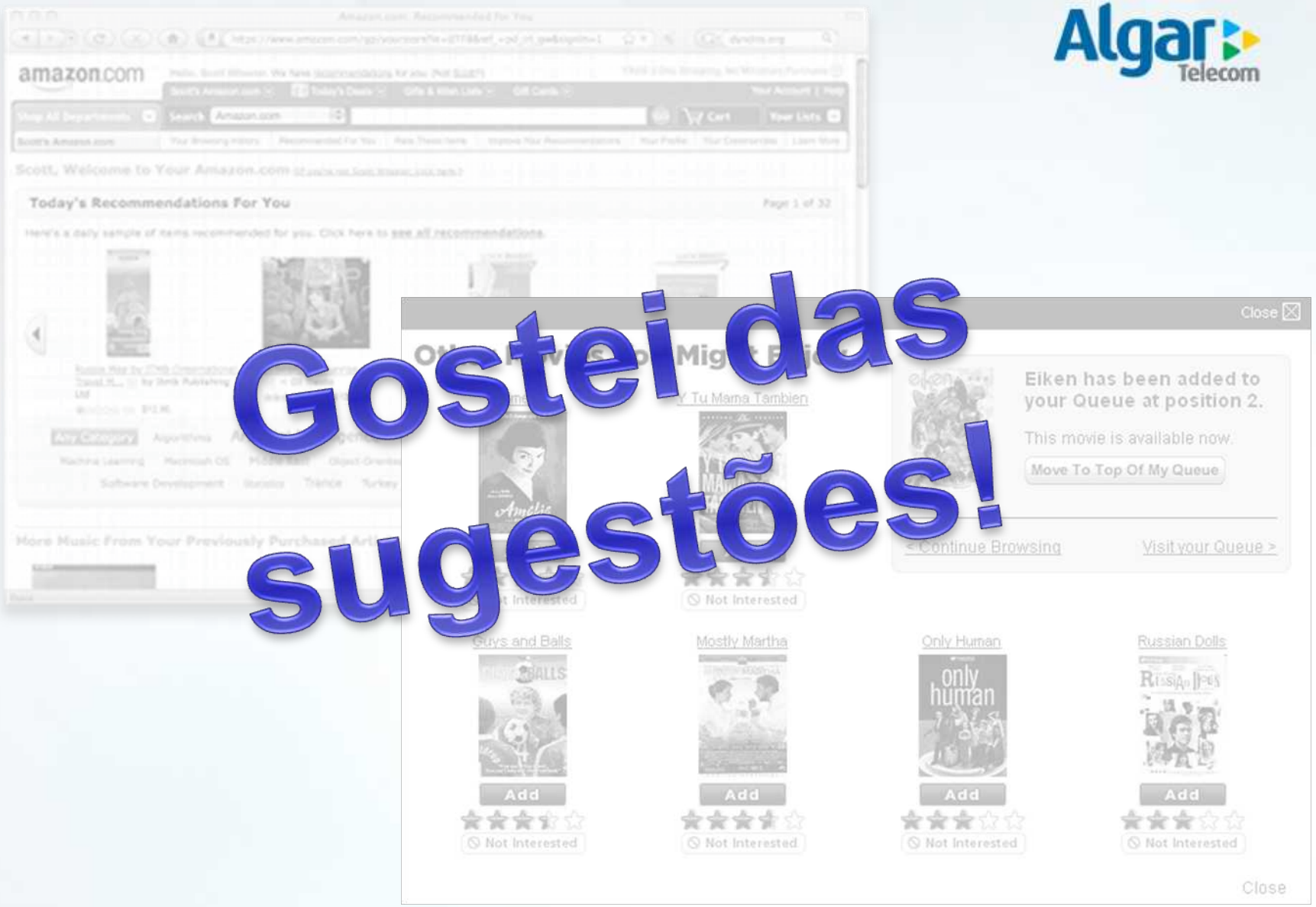
The NY Times (2012)

Science Daily (2009)

Efeito colateral?

Science Daily (2009)


amazon.com
Scott, Welcome to Your Amazon.com
Today's Recommendations For You
Here's a daily sample of items recommended for you. Click here to see all recommendations
Artificial Intelligence
More Music From Your Previously Purchased Artists
Close
Other Movies You Might Enjoy
Amelie
Add
Not Interested
Y Tu Mama Tambien
Add
Not Interested
Eiken has been added to your Queue at position 2.
This movie is available now.
Move To Top Of My Queue
< Continue Browsing
Visit your Queue >
Guys and Balls
Add
Not Interested
Mostly Martha
Add
Not Interested
Only Human
Add
Not Interested
Russian Dolls
Add
Not Interested
Close



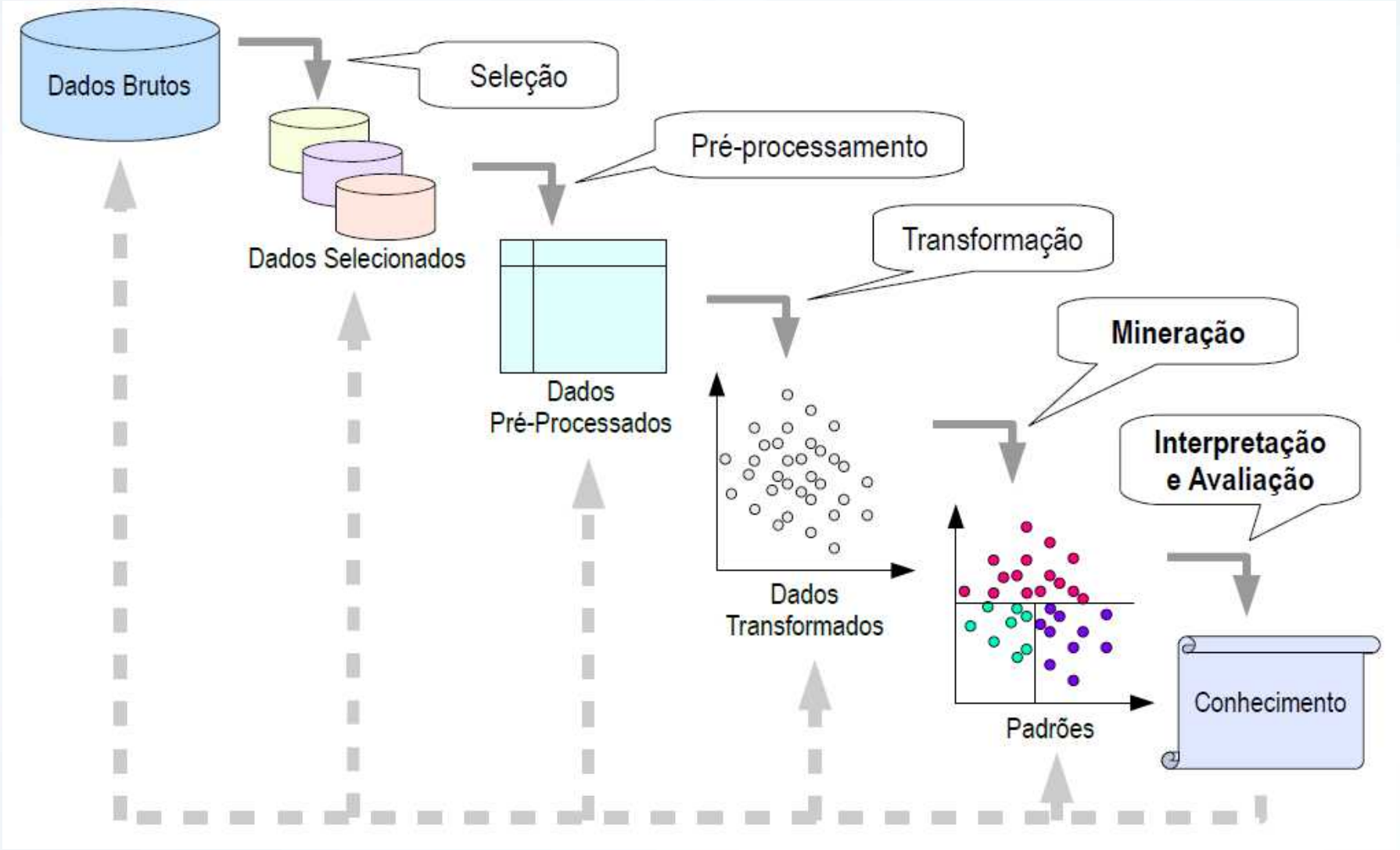
Gostei das sugestões!

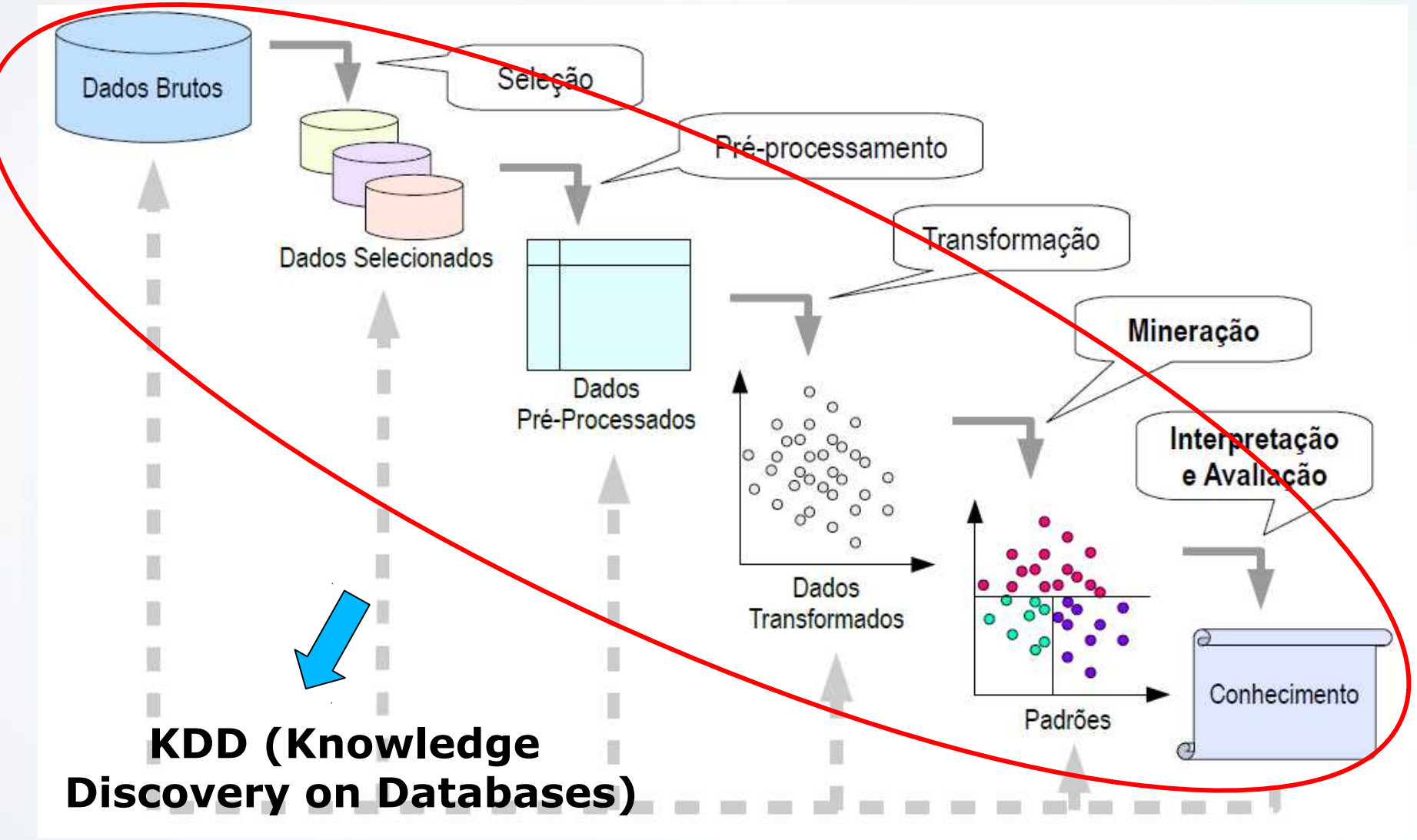
Roteiro


Mineração de Dados e KDD
Tendências de Pesquisa na Área
Apache Mahout e Técnicas de Mineração
PoC
Trabalhos Futuros

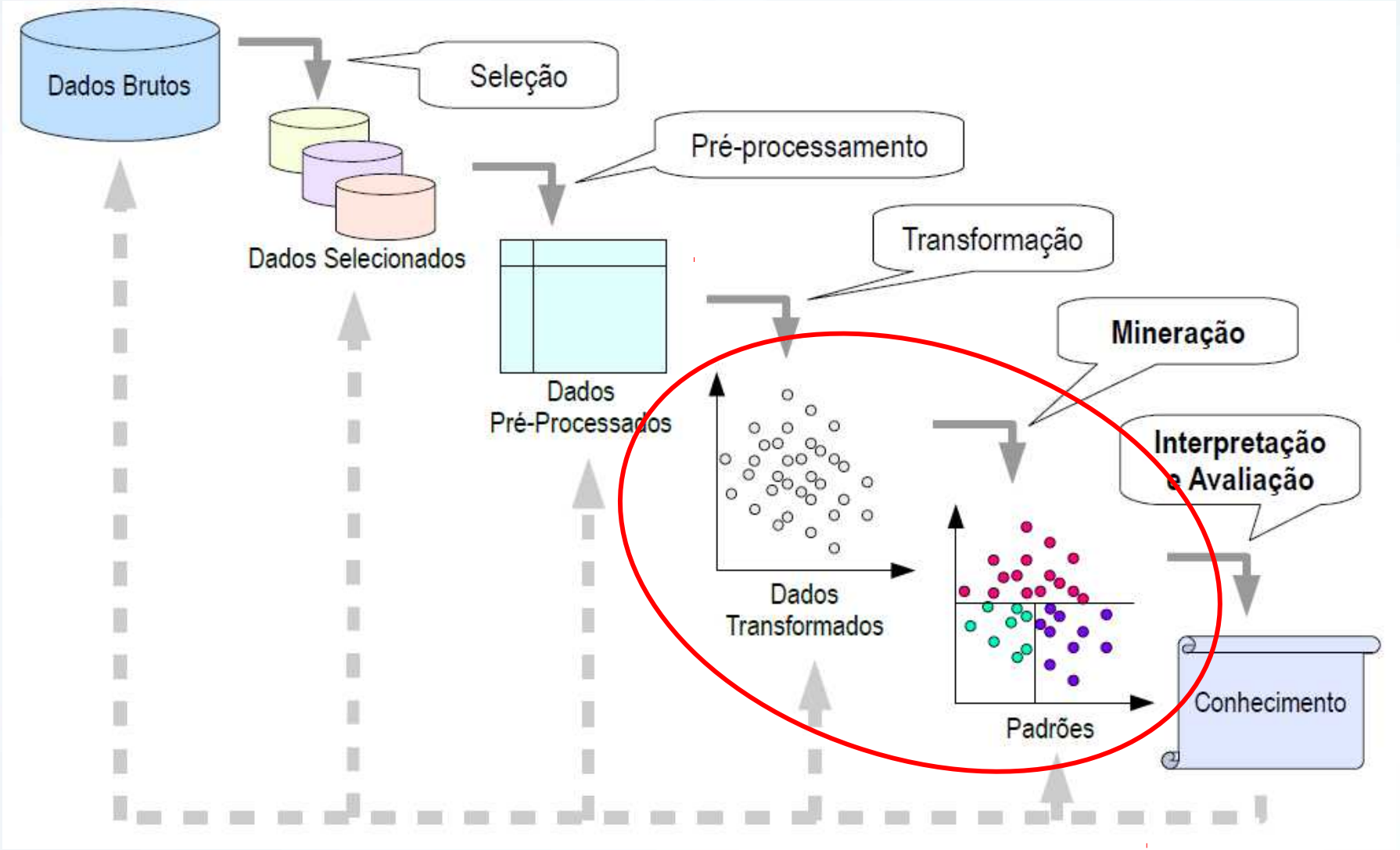
O que é Mineração de Dados?

(em 6 slides)


Dados Brutos
Seleção
Dados Selecionados
Pré-processamento
Dados
Pré-Processados
Transformação
Dados
Transformados
Mineração
Padrões
Interpretação
e Avaliação
Conhecimento

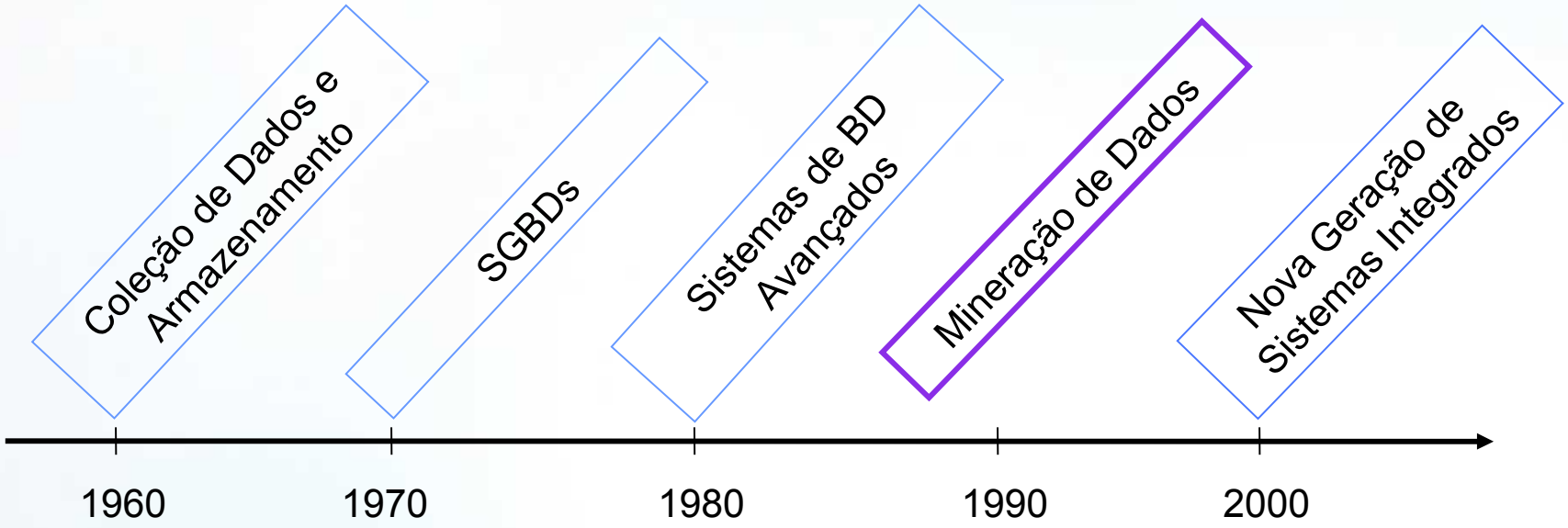


KDD (Knowledge Discovery on Databases)


Dados Brutos
Seleção
Dados Selecionados
Pré-processamento
Dados
Pré-Processados
Transformação
Mineração
Dados
Transformados
Interpretação
e Avaliação
Padrões
Conhecimento

“É a transformação de dados em conhecimento, através da descoberta de padrões”

Histórico


Coleção de Dados e Armazenamento
SGBDs
Sistemas de BD Avançados
Mineração de Dados
Nova Geração de Sistemas Integrados
1960
1970
1980
1990
2000

DM é interdisciplinar

Quais tipos de dados são minerados?

Bancos de Dados Relacionais

Análise dos dados de clientes (idade, salário) para prever o risco de crédito para novos clientes

Exemplo de um BD Relacional

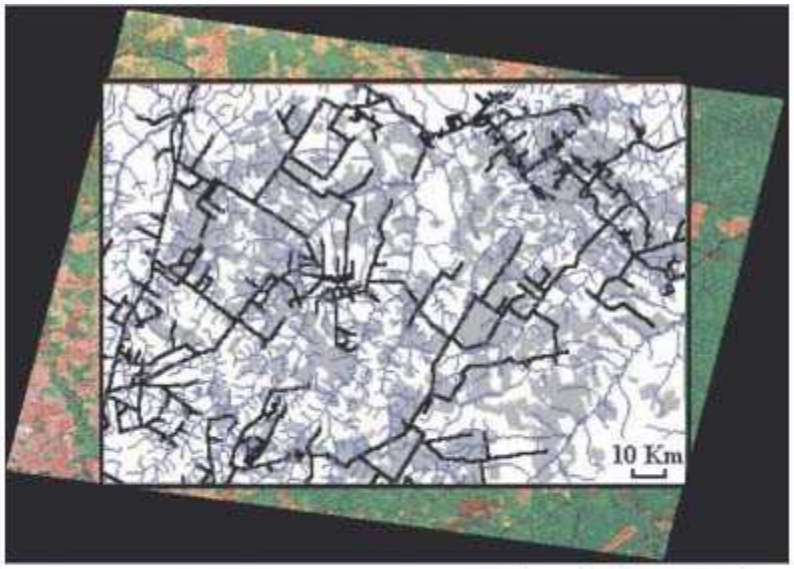
Empregado

NumEmp	NomeEmp	Salário	Dept
032	J Silva	380	21
074	M Reis	400	25
089	C Melo	520	28
092	R Silva	480	25
112	R Pinto	390	21
121	V Simão	905	28
130	J Neves	640	28

Departamento

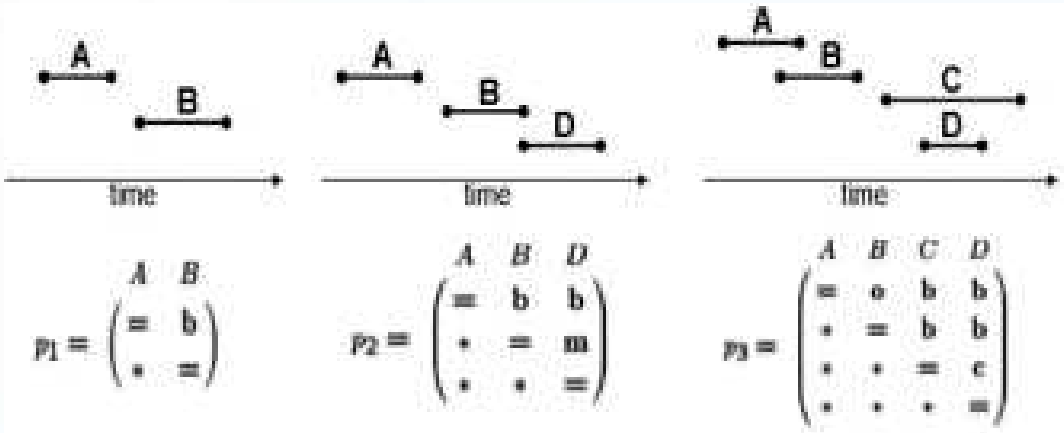
NumDept	NomeDept	Ramal
21	Pessoal	142
25	Financeiro	143
28	Técnico	144

Bancos de Dados Espaciais



Descobrir o comportamento do clima em áreas montanhosas

Bancos de Dados Temporais



$$P_1 = \begin{array}{c} \begin{matrix} A & B \end{matrix} \\ \begin{pmatrix} = & b \\ \cdot & = \end{pmatrix} \end{array}$$

$$P_2 = \begin{array}{c} \begin{matrix} A & B & D \end{matrix} \\ \begin{pmatrix} = & b & b \\ \cdot & = & m \\ \cdot & \cdot & = \end{pmatrix} \end{array}$$

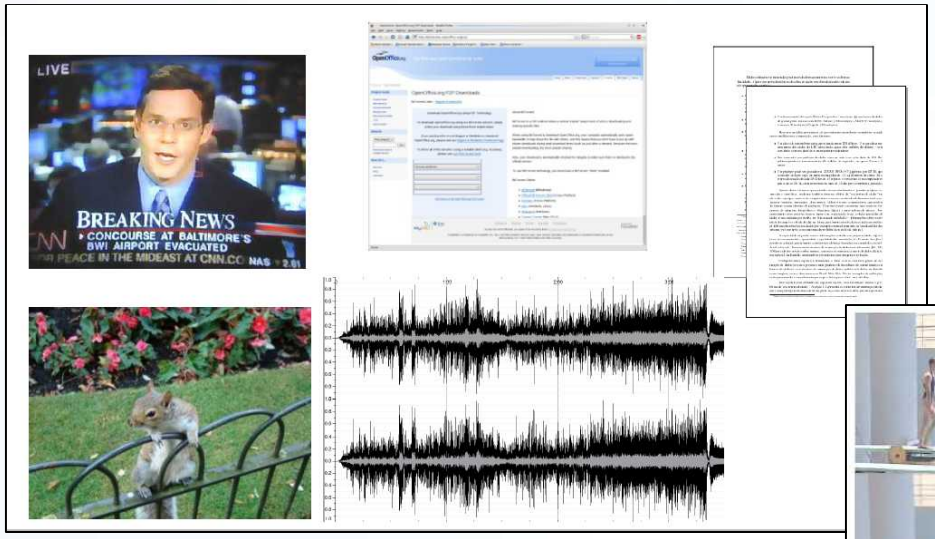
$$P_3 = \begin{array}{c} \begin{matrix} A & B & C & D \end{matrix} \\ \begin{pmatrix} = & o & b & b \\ \cdot & = & b & b \\ \cdot & \cdot & = & c \\ \cdot & \cdot & \cdot & = \end{pmatrix} \end{array}$$

Qual a melhor forma de renovar o estoque? Quais produtos para determinada época do ano?

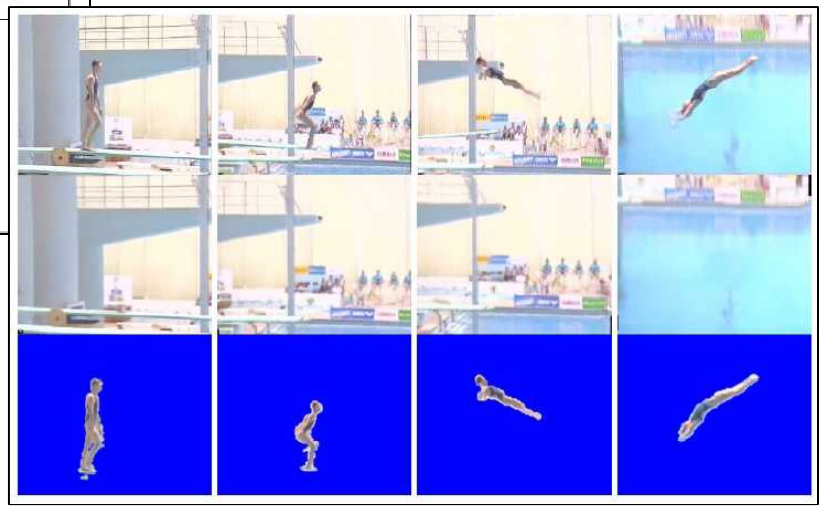
Bancos de Dados de Textos

Minerar especificações, relatórios de erros, tweets, posts, reviews

Bancos de Dados de Áudio, Vídeo e Imagens



Reconhecimento de fala



Segmentação de imagens

Texto, áudio, imagem, relacionais, temporais, espaciais, …

“É a transformação de dados em conhecimento, através da descoberta de padrões”

Anomalias, grupos, classes, tendências, regras, …

CONHECIMENTO

Meu cliente está ansioso com a vinda da concorrência.

O Coreo está sendo utilizado para trotes.

Este cartão de crédito foi clonado.

Roteiro

- Mineração de Dados e KDD
- **Tendências de Pesquisa na Área**
- Apache Mahout e Técnicas de Mineração
- PoC
- Trabalhos Futuros

Streams

Social

Imagens

Web Data Mining

Áudio

Opiniões/Sentimentos

HTML

Vídeos

Streams

Social

Imagens

BIG DATA MINING

Áudio

Opiniões/Sentimentos

HTML

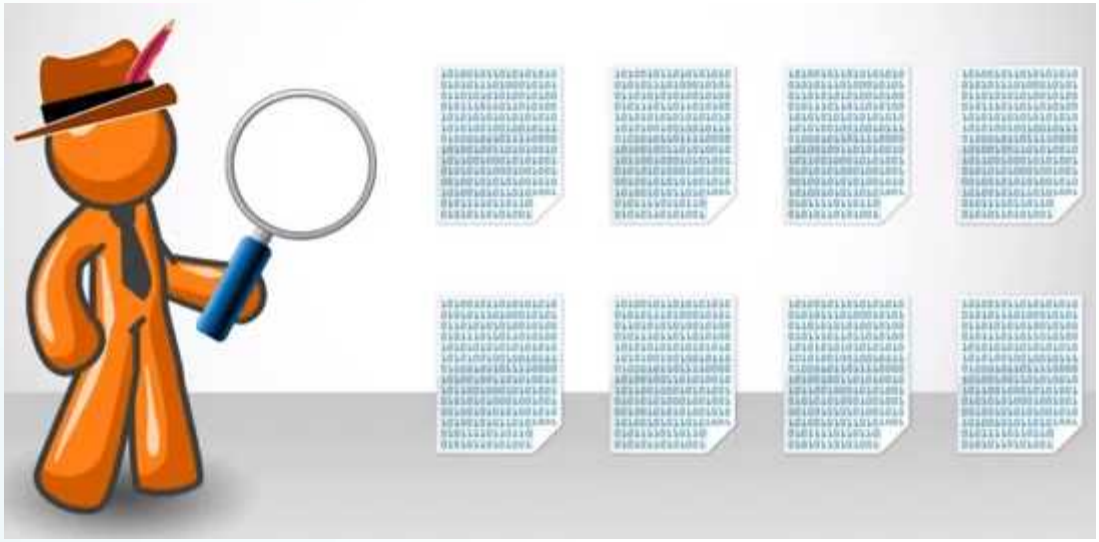
Vídeos

Roteiro

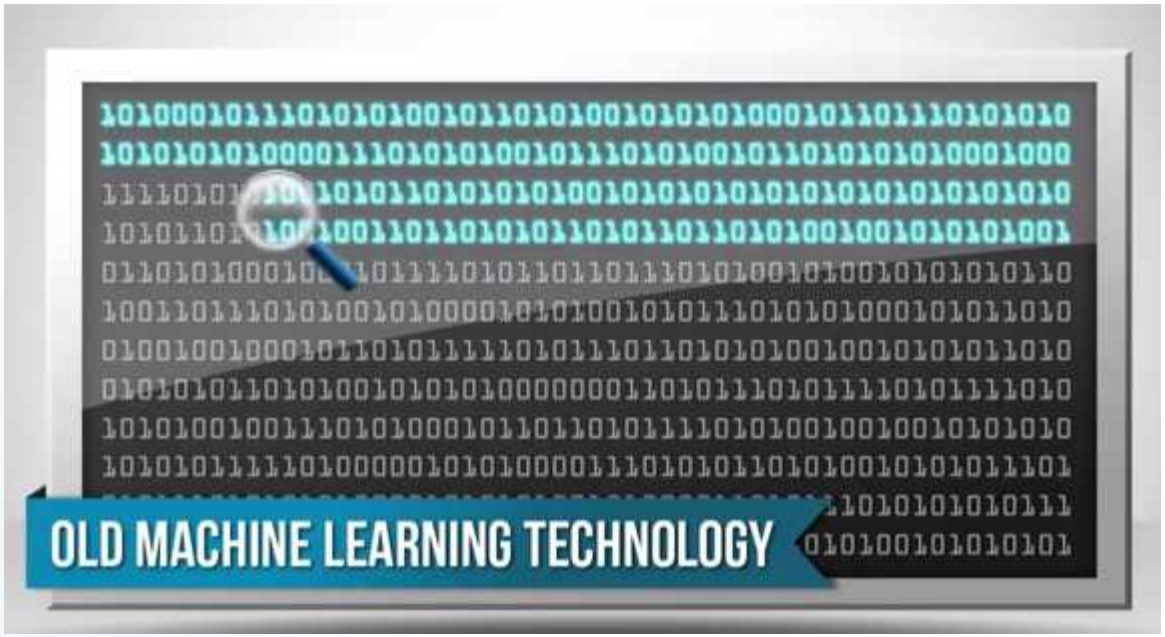

Mineração de Dados e KDD
Tendências de Pesquisa na Área
Apache Mahout e Técnicas de Mineração
PoC
Trabalhos Futuros

O que é o Mahout?

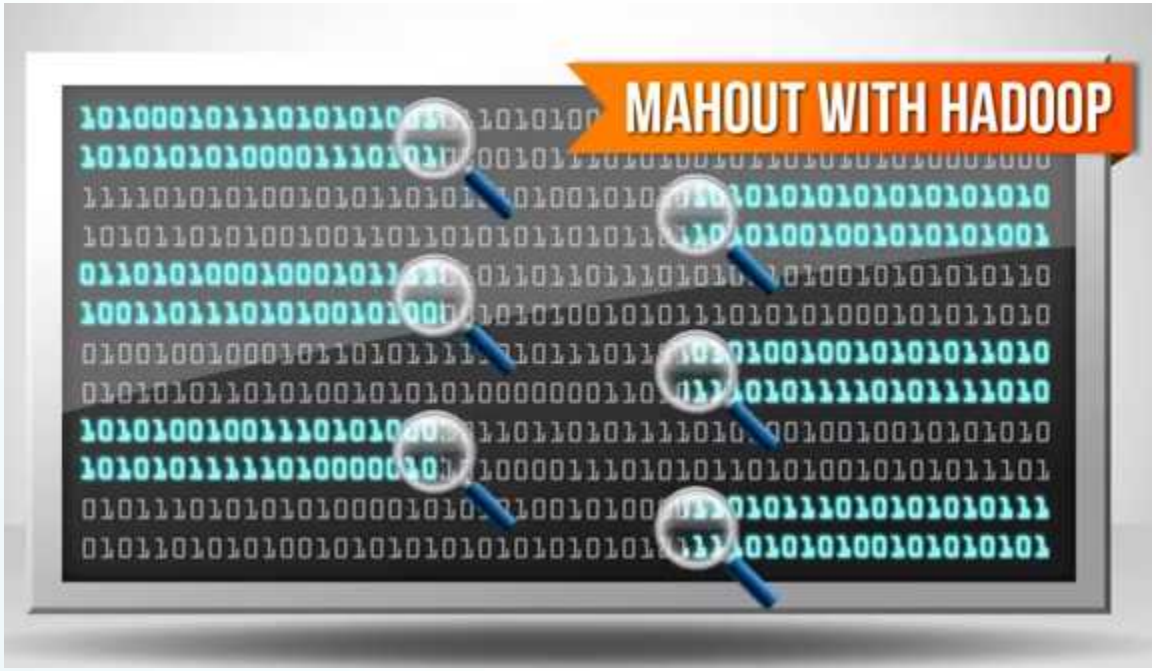
Mahout é um software de "*Machine Learning*" e Mineração de Dados



Mahout é um software de "*Machine Learning*" e Mineração de Dados



Mahout é um software de "*Machine Learning*" e Mineração de Dados



Ficha técnica:

- Projeto Apache
- Open Source
- Última versão: 0.7
- API Java
- Diversos exemplos prontos
- Diversos algoritmos de DM prontos: K-Means, CF, Naïve Bayes, …

Técnicas de Mineração de Dados

- Recomendação
- Clusterização
- Classificação

Recomendação

Recomendação ou Filtro Colaborativo

Recomendação ou Filtro Colaborativo

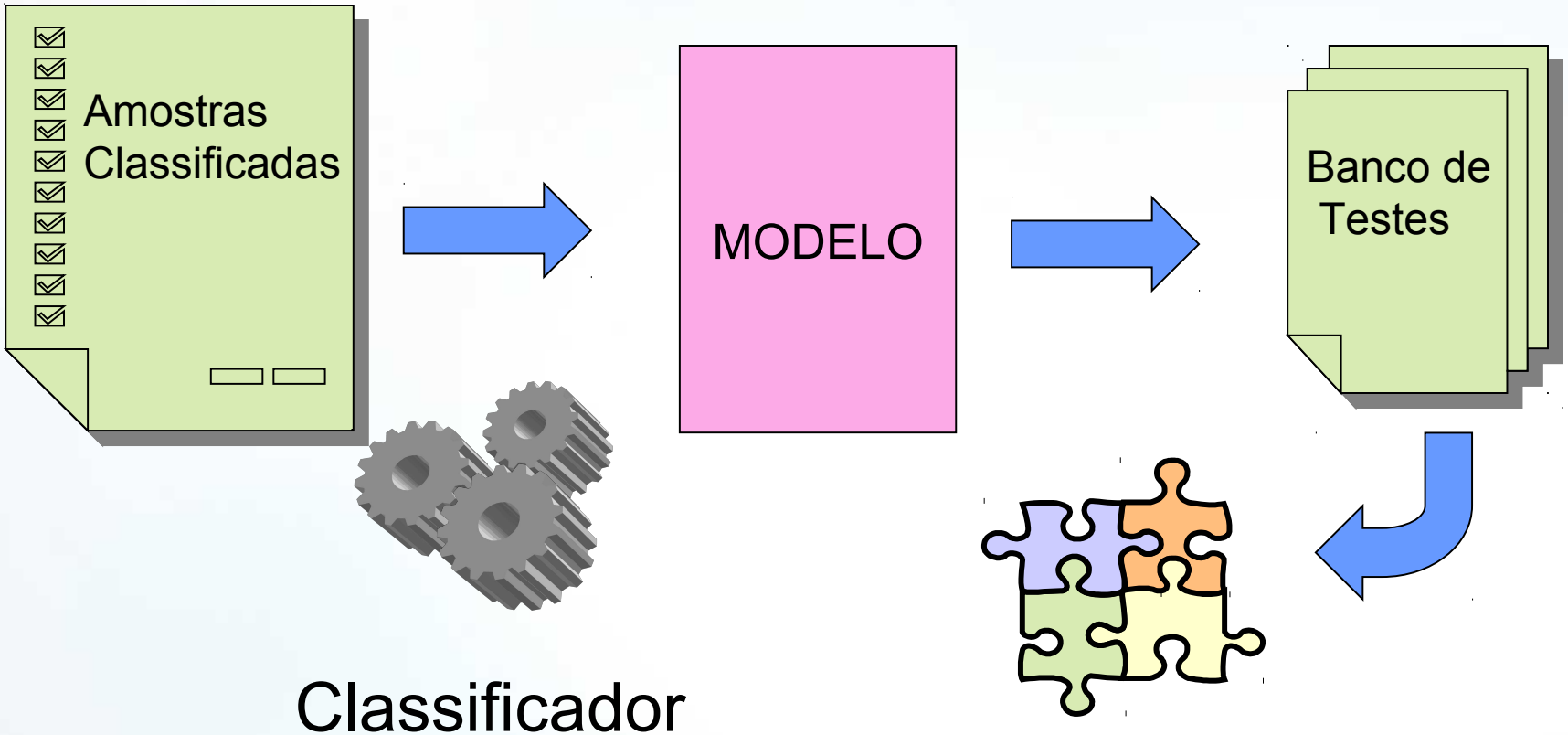
Classificação

Classificação

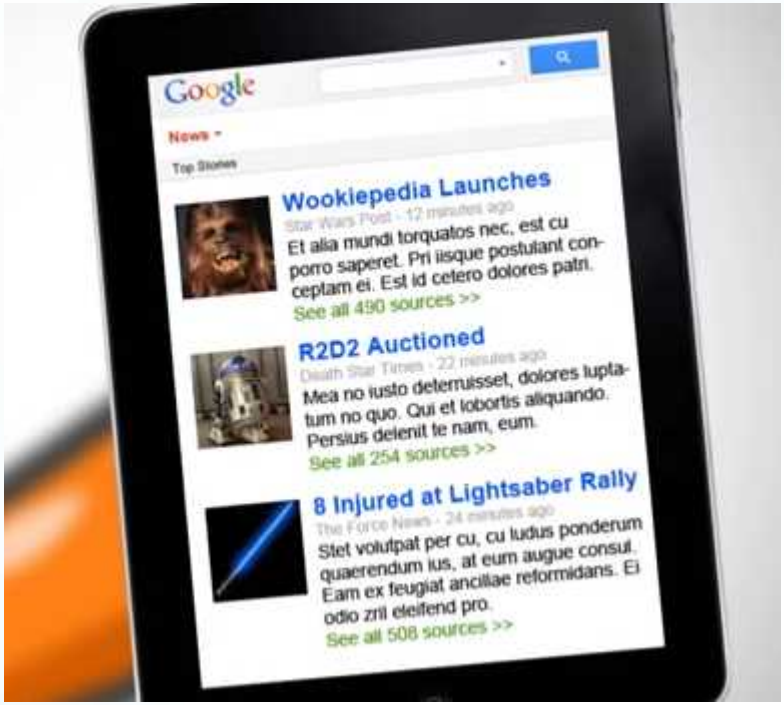
Classificação

Classificação

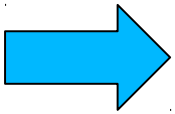
Classificação

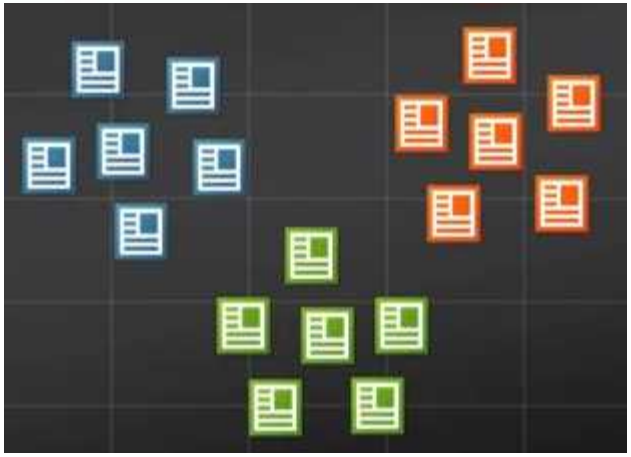


Clusterização ou Agrupamento



Clusterização ou Agrupamento





Clusterização ou Agrupamento

Na prática…

```
$MAHOUT_HOME/bin/mahout wikipediaXMLSplitter -d
$MAHOUT_HOME/examples/temp/enwiki-latest-pages-articles10.x
ml -o wikipedia/chunks -c 64
```

```
$MAHOUT_HOME/bin/mahout trainclassifier -i wikipediainput
-o wikipediamodel
```

```
$MAHOUT_HOME/bin/mahout testclassifier -m wikipediamodel -d
wikipediainput
```

Roteiro


Mineração de Dados e KDD
Tendências de Pesquisa na Área
Apache Mahout e Técnicas de Mineração
PoC
Trabalhos Futuros

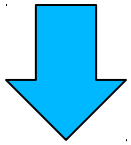
PoC Mahout

- Objetivo: separar textos em SPAM ou NÃO SPAM
- Tipos de dados: texto
- Técnica: classificação
- Algoritmo: Naïve Bayes
- Base: 20news group (spamassassin.apache.org/publiccorpus/20021010_spam.tar.bz2)
- 3050 arquivos

Roteiro


Mineração de Dados e KDD
Tendências de Pesquisa na Área
Apache Mahout e Técnicas de Mineração
PoC
Trabalhos Futuros

Inserir mineração de dados e *machine learning* no BI da empresa



Minerar dados reais

Obrigada