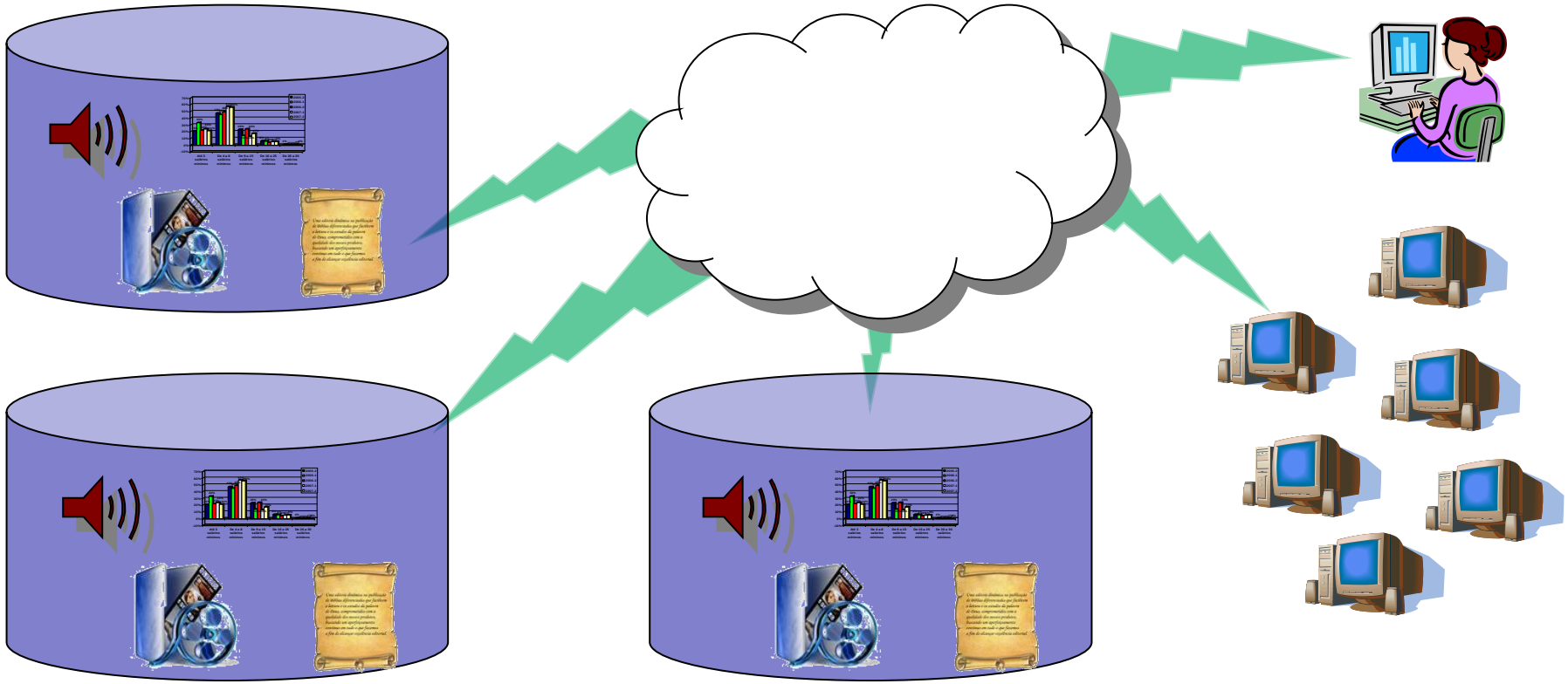
INTRODUÇÃO

BIG DATA

MOTIVAÇÃO

Grandes Desafios da Pesquisa em Computação no Brasil (SBC, 2006)

Gestão da Informação em Grandes Volumes de Dados Multimída Distribuídos



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

POSICIONAMENTO

Grandes Desafios da Pesquisa em Computação no Brasil (SBC, 2014)

Gestão da Informação em Grandes Volumes de Dados Multimída Distribuídos

Ciência de Dados

Astronomia
Biologia
Defesa
Educação
Energia
Engenharia
Esporte
Física
Saúde
Etc...

Computação:
- Gerência de Dados
- Análise de Dados

Temas Relacionados:
- Workflows Científicos
- Procedência de Dados
- Web Semântica
- Mineração de Dados
- Etc...

FUNDAMENTOS E CONCEITOS BÁSICOS

BIG DATA

- Não existe consenso quanto à definição.

- Guarda-Chuva que abriga fundamentos, conceitos e tecnologias voltadas à **gestão** e **análise** de grandes volumes de dados.

- Questões prioritárias no contexto de Big Data: 3Vs / 5Vs
  - Volume
  - Velocidade
  - Variabilidade de forma e conteúdo
  - Valor
  - Veracidade

FUNDAMENTOS E CONCEITOS BÁSICOS

NoSQL

- Assunto fortemente relacionado ao Big Data.
- Também não existe consenso quanto à definição.
- “NoSQL é um conjunto de conceitos e tecnologias relacionados a desempenho, confiabilidade e agilidade que permitam processamento rápido e eficiente de coleções de dados.” (McCreary e Kelly, 2014)
- Provê contraponto aos SGBDRs tradicionalmente encontrados nas empresas durante as últimas décadas.
- Não significa exclusão do uso de recursos de SGBDRs e SQL.

FUNDAMENTOS E CONCEITOS BÁSICOS

NoSQL

- Ambientes NoSQL – Características Principais:
  - Armazenam e recuperam dados em vários formatos.
  - Permitem recuperações de dados sem a realização de junções de estruturas de dados.
  - Permitem a distribuição (com ou sem replicação) de bases de dados em múltiplos processadores que podem ou não estar na nuvem computacional e, apresentar ou não memória compartilhada.
  - Permitem distribuição de processamento, obtendo, em geral, escalabilidade linear em relação ao número de processadores.

- BASE vs ACID: Ambientes NoSQL admitem inconsistência temporária de dados em prol de sua alta disponibilidade (24 x 7)

- Para vantagens e desvantagens de NoSQL, vide: Mohan (2013) e Wayne (2012)

FUNDAMENTOS E CONCEITOS BÁSICOS

NoSQL – Padrões Arquiteturais de Dados

- Pares do Tipo Chave-Valor (Key-Value Stores)
  - Cadeia de símbolos (chave) leva a um blob de dados arbitrariamente grande (valor)

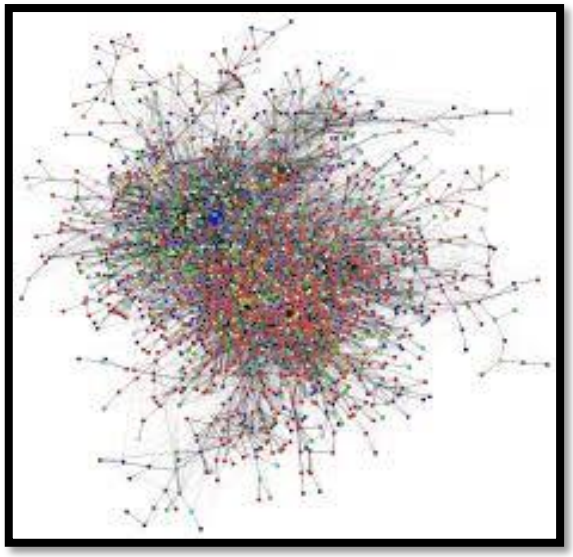
Chave	Valor
Imagem123.jpg	Arquivo binário contendo a imagem
www.ime.eb.br	HTML de uma página web
C:/Documentos/LivroKDD.pdf	Documento PDF

  - Não possuem linguagens de consulta específicas
  - São indexados por chaves que permitem o acesso direto aos dados
  - Valores podem ser de qualquer tipo de dados
  - Vantagem: simplicidade de estrutura: economia de tempo e recursos
  - Exemplos de ambientes que utilizam este padrão:
    - Cassandra
    - Dynamo
    - Voldemort
    - Riak

FUNDAMENTOS E CONCEITOS BÁSICOS

NoSQL – Padrões Arquiteturais de Dados

- Bases de Dados de Grafos (Graph Stores)
  - Armazenamento e recuperação de informações em grafos
  - Dados são triplas: vértice-relacionamento-vértice
  - Vértices e relacionamentos podem conter propriedades
  - Informações podem estar contidas em vértices ou na estrutura dos grafos
  - Forte aplicação: modelagem de redes complexas como as redes sociais, por ex.
  - Oferecem linguagens de consulta específicas. Ex: Cypher (Neo4j)
  - Exemplos de BDs de grafos:
    - Neo4j
    - Allegro Graph
    - DEX
    - Infinite Graph



FUNDAMENTOS E CONCEITOS BÁSICOS

NoSQL – Padrões Arquiteturais de Dados

- Bigtables (Column-Oriented Stores)
  - Admitem chaves complexas, formada por duas ou mais informações.

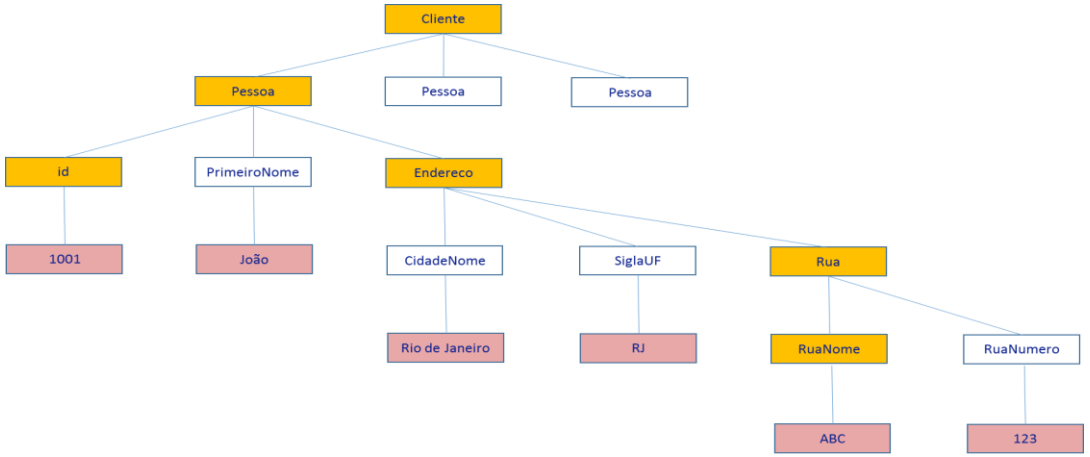
Row-ID	ColumnFamily	ColumnName	TimeStamp	Valor
←----------→				

  - Campo *ColumnFamily* agrupa as colunas indicadas em *ColumnName*.
  - *Timestamp* associa aspecto temporal às informações (versionamento dos dados)
  - Permite armazenar matrizes esparsas de alta dimensionalidade
  - Também dispensa o uso da operação de junção
  - Exemplos de ambientes que utilizam este padrão:
    - Cassandra
    - HBase
    - Hypertable

FUNDAMENTOS E CONCEITOS BÁSICOS

NoSQL – Padrões Arquiteturais de Dados

- Coleções de Documentos (Document Stores)
  - Cada documento corresponde a uma árvore.



  - Conteúdo da árvore pode ser acessado via linguagem apropriada (ex: SPARQL)
  - *JSON* e *XML* são exemplos de formatos de apresentação de documentos
  - Exemplos de sistemas de gestão de documentos:
    - MongoDB
    - RavenDB
    - CouchDB

FUNDAMENTOS E CONCEITOS BÁSICOS

MapReduce

- Modelo de programação: processa grandes volumes de dados em paralelo
  - "Proposto" pela Google em 2004.
  - Exemplo que ilustra o funcionamento do MapReduce:
    - Distribuição da tarefa de contagem da população de Roma por regiões.
    - A contagem em cada conjunto ocorre em paralelo.
    - Dados levantados são consolidados no número final
  - Outro exemplo:

No fim de cada iteração, é realizada a soma das frequências locais de cada conjunto de itens, o que resulta na frequência global de cada um deles.

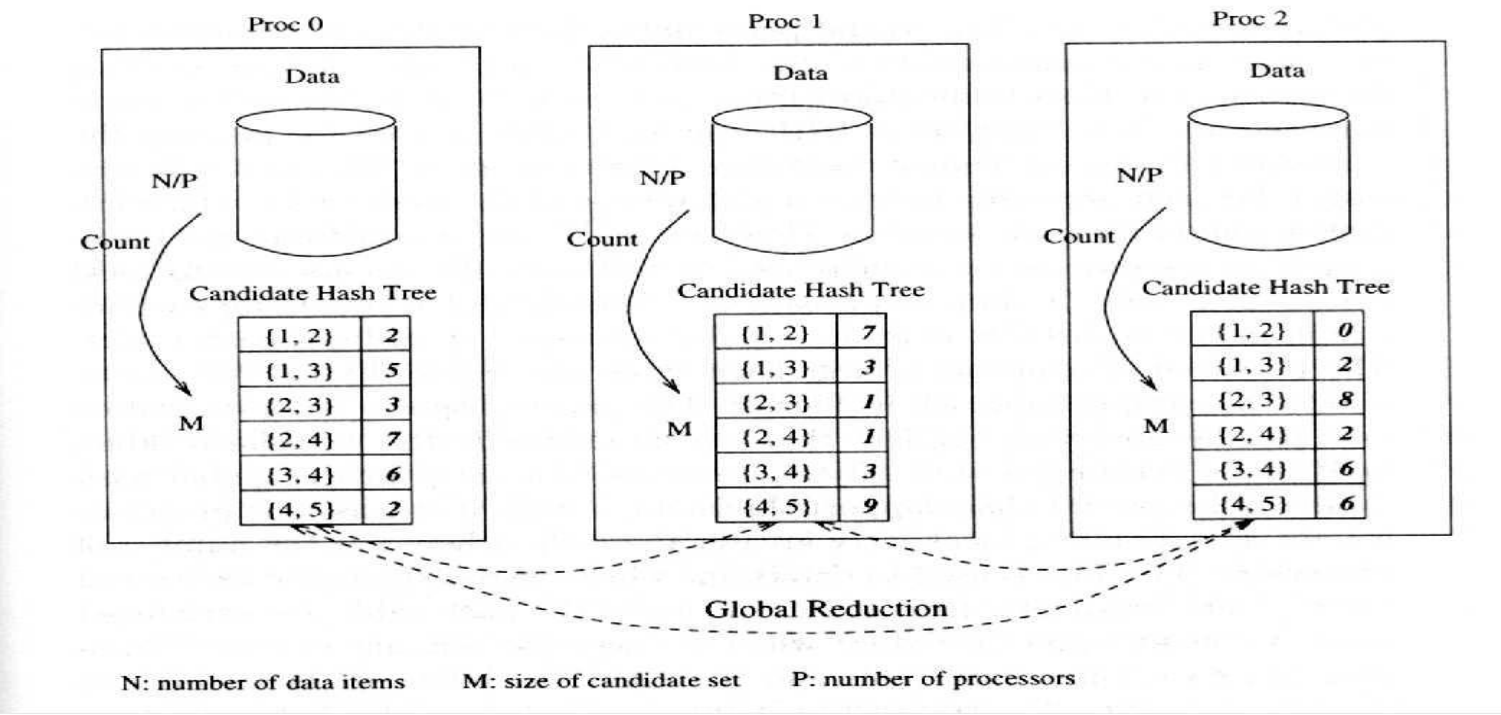
Chave	Valor
1	[a, é, o]
2	[No, de, de, de, de, na, um]
3	[das, que, fim]
4	[soma, cada, cada, cada]
5	[itens, deles]
6	[locais, global]
7	[resulta]
8	[iteração, conjunto]
9	[realizada]
10	[frequência]
11	[frequências]

Chave	Valor
1	3
2	7
3	3
4	4
5	2
6	2
7	1
8	2
9	1
10	1
11	1

FUNDAMENTOS E CONCEITOS BÁSICOS

Hadoop

- Projeto da Fundação Apache
  - Oferece framework para operações paralelas em grandes volumes de dados.
  - Funciona sobre sistema de arquivos organizados em clusters distribuídos.
  - Se baseia no paradigma do MapReduce.



FUNDAMENTOS E CONCEITOS BÁSICOS

Análise de Dados

Necessidade:

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

Padrões

Tendências

Associações

Mineração de Dados

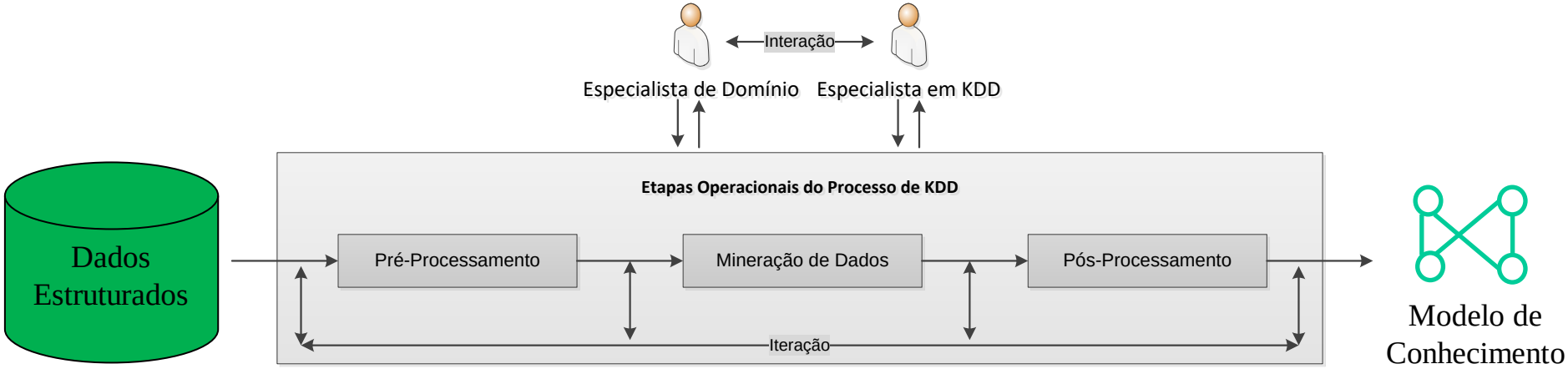
(Data Mining)

Descoberta de Conhecimento em Bases de Dados (KDD)

FUNDAMENTOS E CONCEITOS BÁSICOS

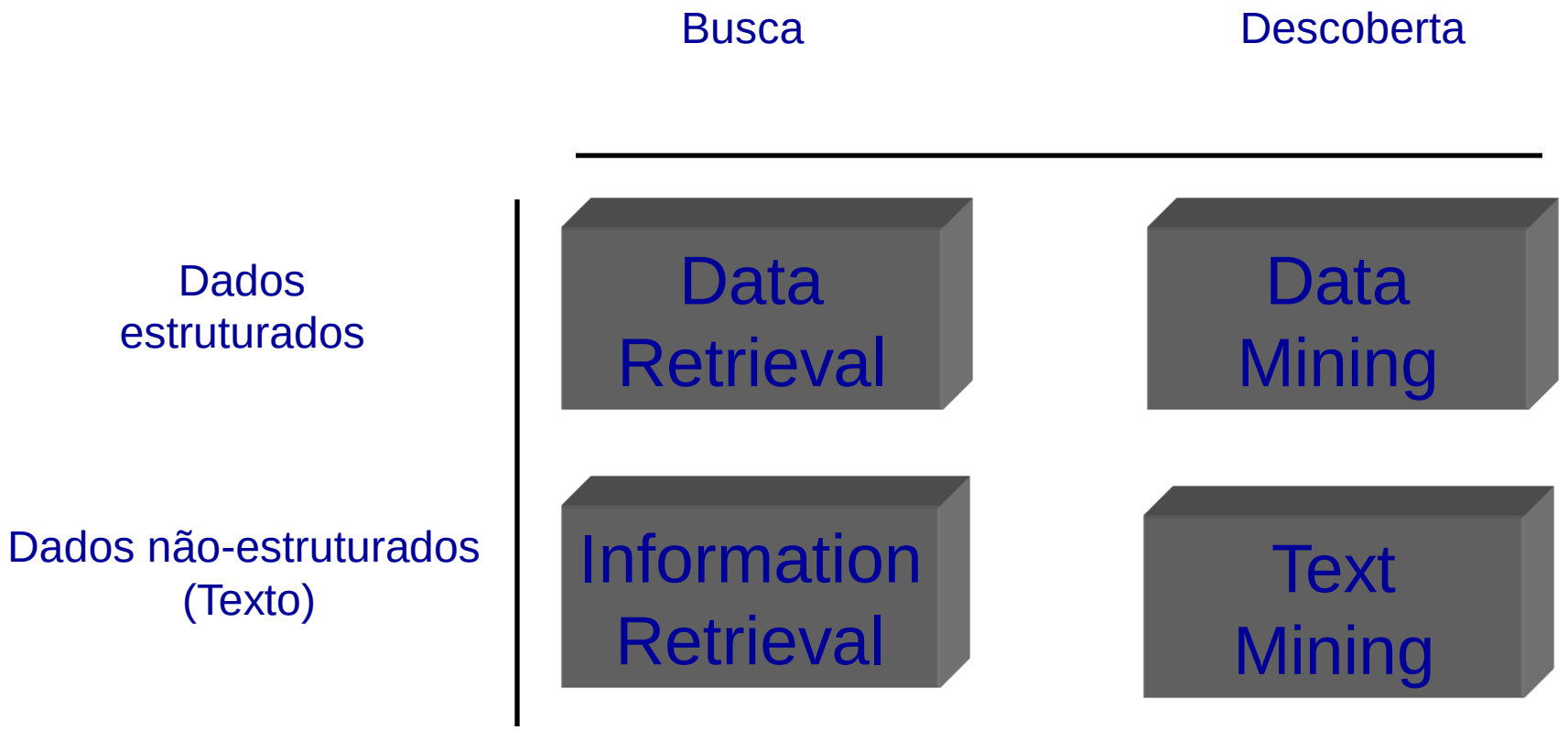
Descoberta de Conhecimento em Bases de Dados – KDD

"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)**



FUNDAMENTOS E CONCEITOS BÁSICOS

"BUSCA" VS "DESCOBERTA"



FUNDAMENTOS E CONCEITOS BÁSICOS

- Há vários tipos de “mining” :
  - Data Mining
  - Multimídia Mining (Som, Imagem, …)
  - Text Mining
  - Graph Mining
  - Web Mining
  - Educational Data Mining (EDM)
  - Social Data Mining
  - Opinion Mining
  - ...

- Terminologia acima não é um consenso.

FUNDAMENTOS E CONCEITOS BÁSICOS

DESCOBERTA DE CONHECIMENTO - UMA TAXONOMIA

Descoberta de Conhecimento

Dados Estruturados	Dados Textuais	Dados Multimídia
Mineração de Dados Clássica	Mineração de Dados Textuais	Mineração de Dados Multimídia