Mineração de Dados

Eduardo Raul Hruschka

Baseado no curso de Gregory Piatetsky-Shapiro, disponível no sítio http://www.kdnuggets.com

Visão Geral:

- Introdução: motivação, aplicações, conceitos básicos.
- Agrupamento de dados (*clustering*).
- Classificação.
- Regras de Associação.

Introdução

- Contextualizar a Mineração de Dados;
- Aplicações de Mineração de Dados;
- Conceitos Básicos;
- Tarefas de Mineração de Dados (MD).

Tendências que nos levam a um cenário de super-abundância de dados:

- Instituições financeiras, telecomunicações, transações em empresas…
- Dados científicos: astronomia, biologia, etc.
- Dados na Web, Dados em textos, comércio eletrônico, …

→ Capacidades de coletar/armazenar superaram nossas habilidades de analisar/extrair conhecimento dos dados:

→ É necessária a aplicação de técnicas/ferramentas que transformem, de maneira inteligente e automática, os dados disponíveis em informações úteis, que representem conhecimento.

Alguns Exemplos:

- AT&T manipula bilhões de chamadas por dia. Todos os dados não podem ser armazenados (~ 26 TB) – análise "on the fly" / "streaming data";
- France Telecom: banco de dados ~30TB;
- Google: procura aproximadamente 4 bilhões de páginas – centenas de TB;
- Wal-Mart: 20 milhões de transações por dia;
- Nasa: 50Gb de imagens por hora;
- Estimativa da UC Berkeley em 2003: 5 hexabytes (5 milhões de terabytes) de dados novos foram gerados em 2002.
  - www.sims.berkeley.edu/research/projects/how-much-info-2003/

Taxa de crescimento dos dados

- A quantidade de informações geradas em 2002 é o dobro do que foi gerado em 1999 (crescimento anual de ~30%);
- Um humano poderia analisar uma quantidade tão grande de dados?
- Descoberta de Conhecimento - Mineração de Dados;
- Analogia: explorar uma mina de dados, purificando-se o minério para obter o ouro (conhecimento).

Introdução

- Contextualizar MD
- **Aplicações de MD**
- Conceitos Básicos
- Tarefas de Mineração de Dados

Aplicações de Mineração de Dados

Há uma certa dificuldade em relatar casos práticos, pois o produto da mineração de dados geralmente oferece alguma vantagem competitiva e, por isso, raramente é descrito...

- Ciência:
  - Astronomia, Bioinformática, Meteorologia,...
- Negócios
  - Propaganda, CRM (Customer Relationship management), Investimentos, Marketing, Planos de Saúde e de Seguros...
- Web:
  - Ferramentas de busca, padrões de navegação...
- Governo
  - Dados demográficos, anti-terror, ...

Mineração de Dados para Modelagem de Clientes:

- Previsão de atritos;
- Marketing direto;
- Avaliação de Risco;
- Detecção de fraudes;

Aplicação para casos de atrito com clientes:

- Situação: taxa de atrito para telefone celulares está em torno de 25-30% por ano (http://www.kdnuggets.com).

Tarefa:

- Considerando-se as informações do cliente para os últimos meses, prever quem entrará em atrito no próximo mês.
- Além disso, estimar o valor do cliente e qual é o custo de uma oferta a ser feita ao cliente para não perdê-lo.

Avaliação de risco para crédito:

- Para quem você emprestaria R$ 500,00 ?

- Situação: pessoa solicita um empréstimo.
- Tarefa: o banco deve aprovar o empréstimo?
- Observação: pessoas com *muito crédito* em geral não precisam de empréstimos, enquanto que pessoas com *pouco crédito* provavelmente não pagarão o empréstimo. Como identificar os melhores clientes?
- Atualmente muitas instituições financeiras adotam técnicas de mineração de dados (classificação) pra resolver este problema.

Comércio eletrônico

- Consideremos que uma pessoa compra um livro na Amazon.com.
- Tarefa: recomendar outros livros (produtos) potencialmente interessantes para a pessoa em questão.
- Amazon utiliza técnicas de agrupamento que podem proporcionar padrões do tipo:
  - Clientes que compram "**Advances in Knowledge Discovery and Data Mining**", também compram "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"

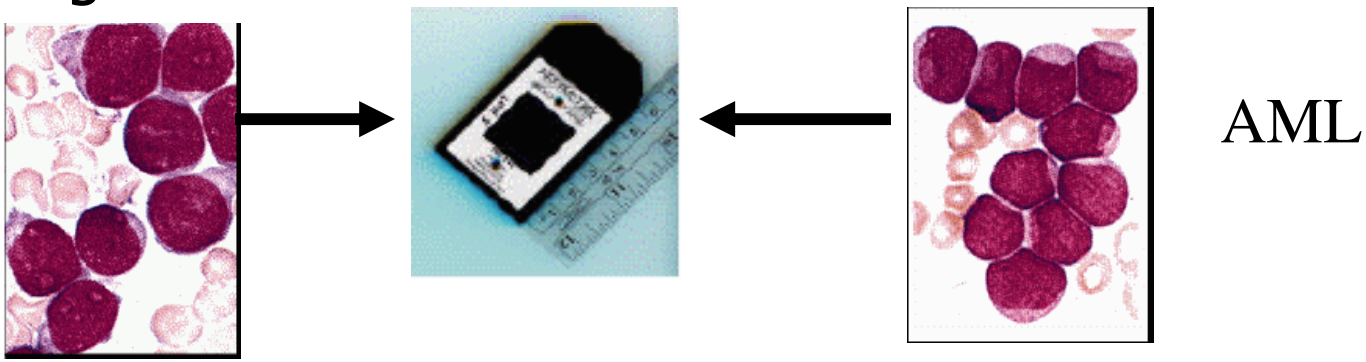
Genomic Microarrays

Considerando-se dados de *microarray* para uma determinada quantidade de pacientes (amostras), é possível:

- Precisamente diagnosticar a doença?
- Prever o resultado para determinado tratamento?
- Recomendar o melhor tratamento?

Exemplo:

- 38 casos de treinamento, 34 teste, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Usar dados de treinamento para construir um modelo para diagnóstico:

ALL  AML

Resultados na base de teste: 33/34=97% corretos.

Problemas adequados para MD:

- Requerem decisões baseadas em *conhecimento;*
- Ambiente dinâmico (dados novos);
- Existem métodos sub-ótimos;
- Há dados acessíveis, relevantes e em quantidade suficiente;
- Proporcionam recompensas elevadas pelas decisões corretas;
- Privacidade é um assunto importante.

Introdução

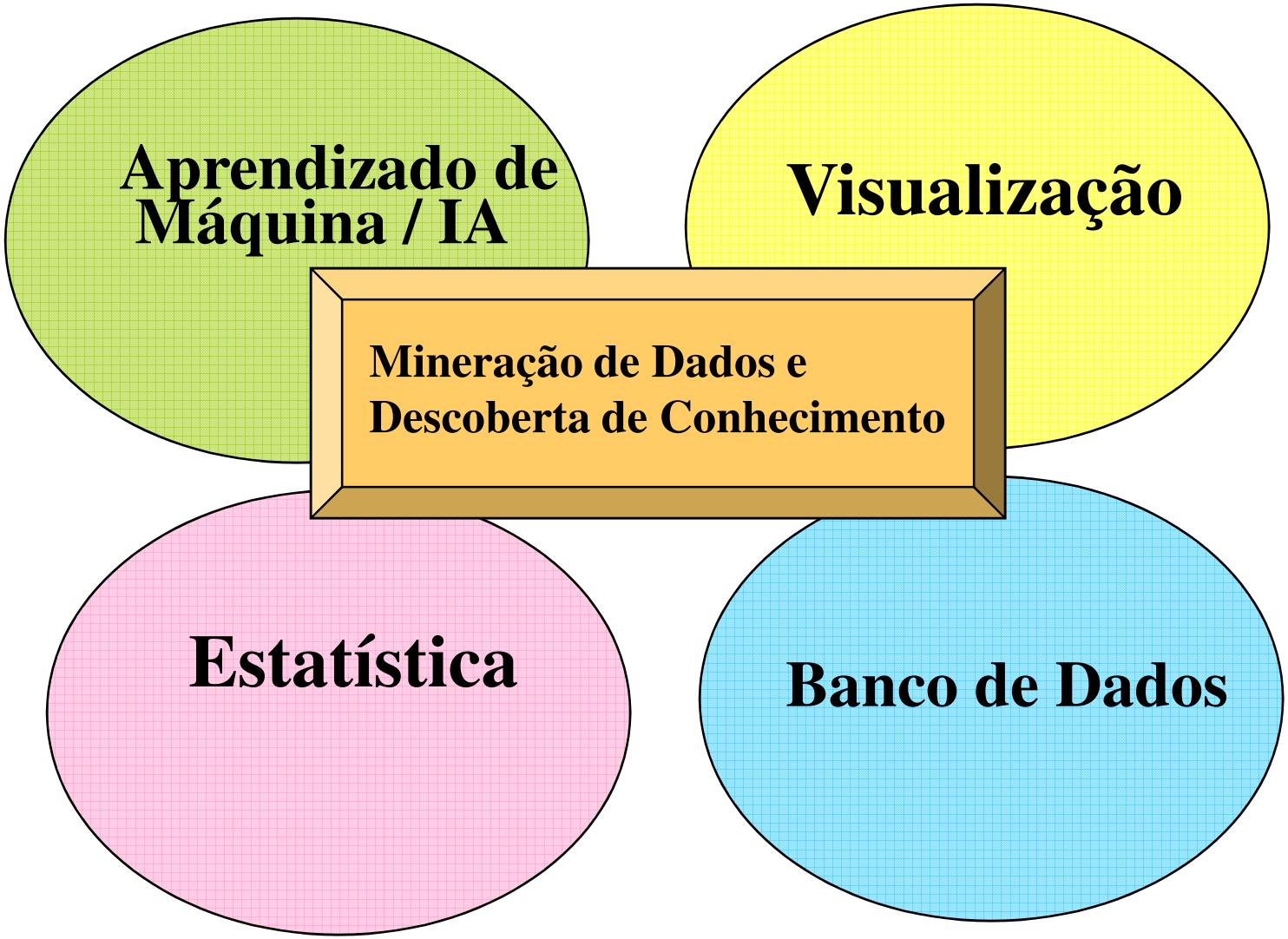
- Contextualizar a Mineração de Dados
- Aplicações de Mineração de Dados
- **Conceitos Básicos**
- Tarefas de Mineração de Dados

Uma definição de KDD (*Knowledge Discovery from Databases*):

KDD é o processo *não trivial* de identificar padrões válidos, novos, potencialmente úteis e compreensíveis em dados.

- *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996.

Áreas correlatas:

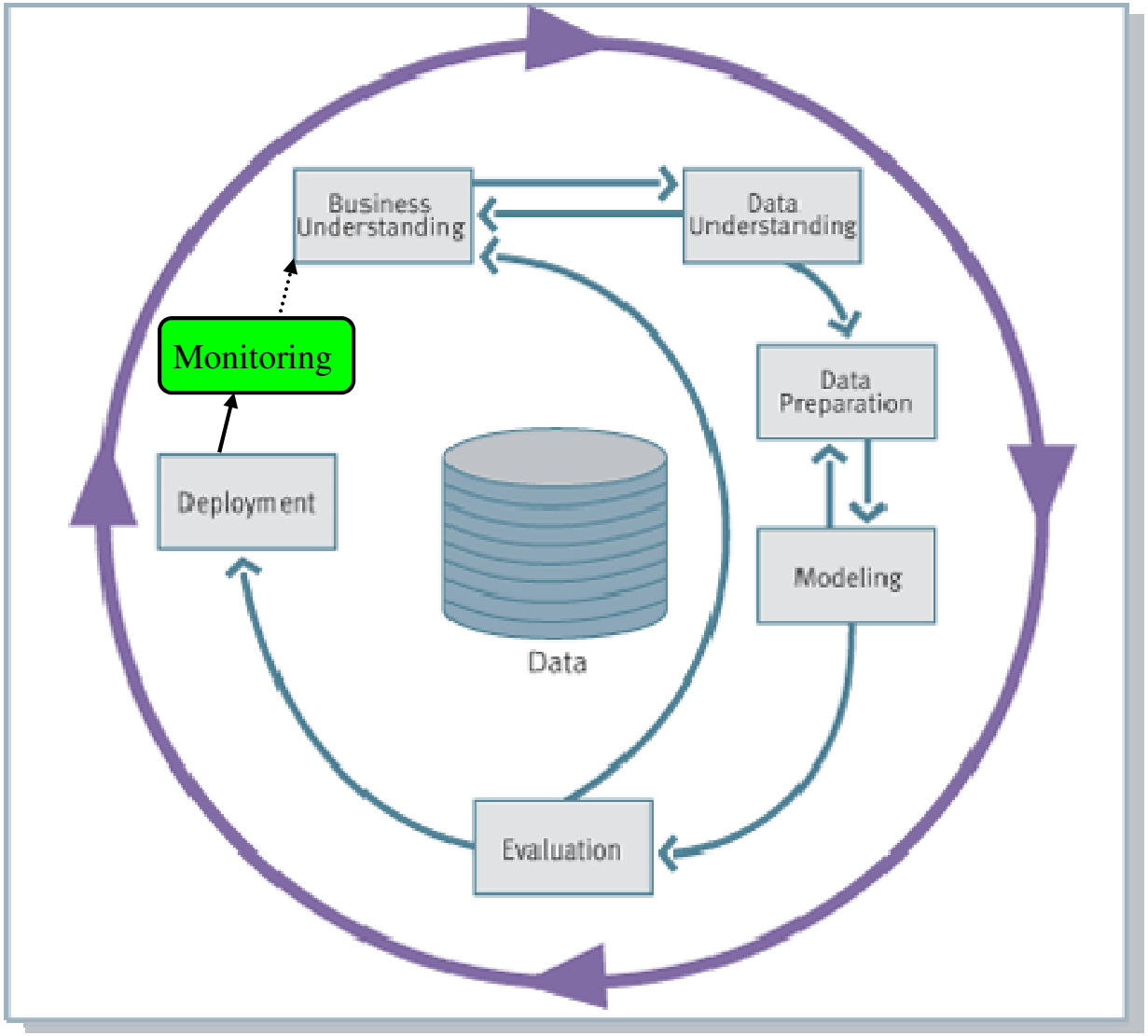


Estatística, Aprendizado de Máquina e Mineração de Dados:

- Estatística:
  - *Mais* baseada na teoria;
  - Mais focada em testes de hipóteses.
- Aprendizado de Máquina:
  - *Mais* heurísticas;
  - Focada em melhorar o desempenho de agentes de aprendizado;
  - Aprendizado em tempo-real e robótica (em geral não abordados em MD).
- MD/KDD – muito empregados indistintamente:
  - Procura integrar teoria e heurísticas;
  - Inclui preparação de dados, visualização de resultados, pós-processamento;
  - Geralmente empregada em *grandes* bases de dados.

* Distinções são mais ou menos *nebulosas!*

Fluxo do KDD:



see www.crisp-dm.org for more information

Introdução

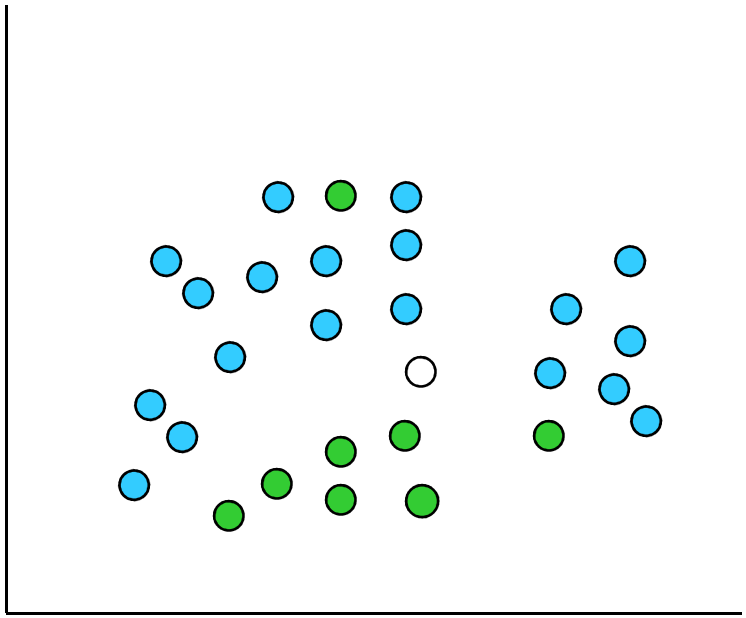
- Contextualizar a Mineração de Dados
- Aplicações de Mineração de Dados
- Conceitos Básicos
- **Tarefas de Mineração de Dados**

Principais tarefas de MD:

- **Classificação:** prever uma classe;
- **Agrupamento:** encontrar clusters;
- **Associações:** e.g. A & B & C ocorrem freqüentemente;
- **Visualização:** facilitar descoberta por humanos;
- **Sumarização:** descrever um grupo;
- **Detecção de Desvios**: encontrar mudanças;
- **Estimação:** prever valores contínuos;
- …

Classificação:

- Aprender um método para prever a classe a partir de exemplos pré-classificados.

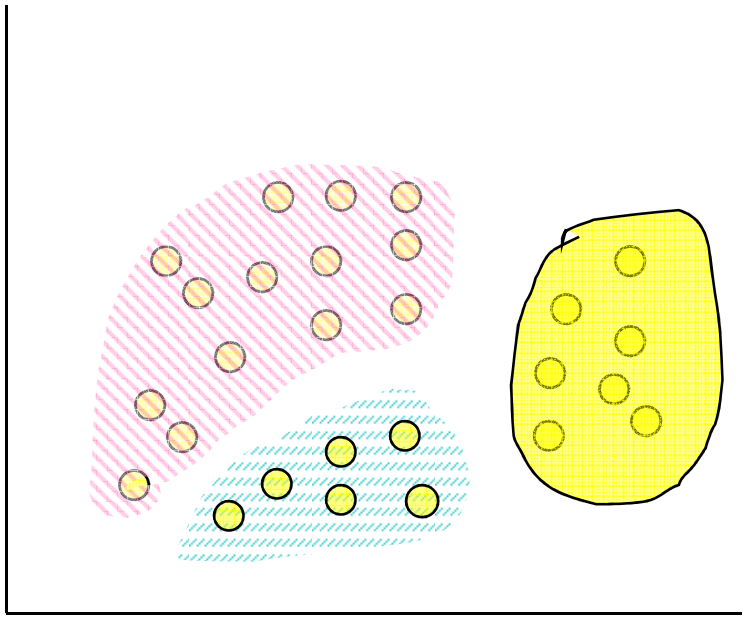


Muitas abordagens:

Estatística,
Árvores de decisão,
Redes neurais,
...

Agrupamento (Clustering):

- Encontrar grupos naturais de exemplos (dados não rotulados)



Sumário:

- Tendências tecnológicas têm nos levado a um problema de super-abundância de dados;
  - MD é necessária para "dar sentido" aos dados;
- MD possui muitas aplicações (importante lembrar que *com* e *sem* sucesso);
- Tarefas de MD:
  - classificação, *clustering*, associação, ...

Mais sobre MD e KDD:

KDnuggets.com

- Notícias, Publicações;
- Software, Soluções;
- Cursos, Encontros, Conferências;
- Publicações, *Websites*, Bases de dados;
- Oportunidades de Empregos
- ...

Alguns Centros de Pesquisa:

- Microsoft Research Group (Cambridge, Beijing, San Francisco);
- IBM Research Center (8 centros) / Intelligent Miner;
- HP Labs – Cambridge, Palo Alto.