Mineração de Dados

Introdução

Sumário

Disciplina

Informações Gerais

- Professor
  - Heder – heder@ice.ufjf.br
- Horário
  - Quarta-feira, 19h-21h
  - Sexta-feira, 21h-23h

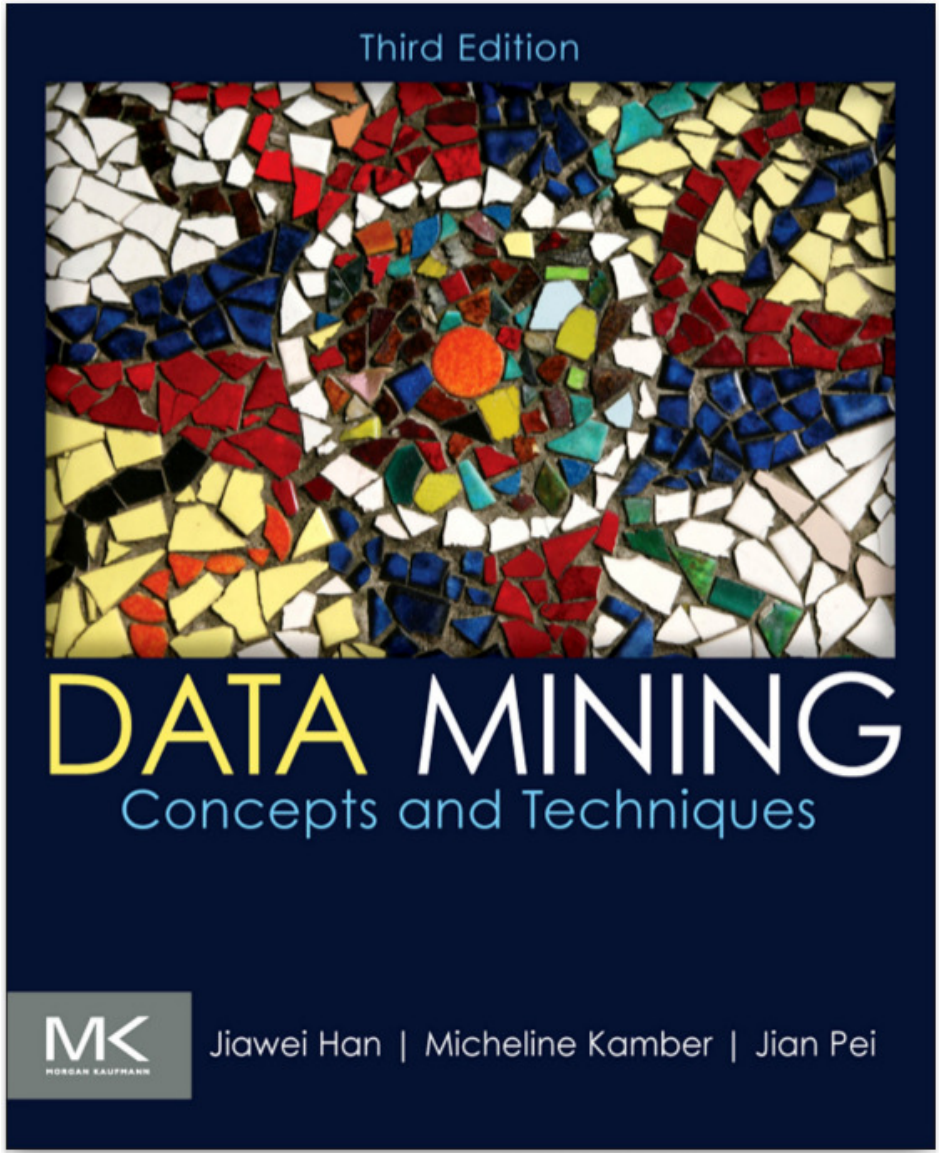
Conteúdo

- Descoberta de conhecimentos em bases de dados
- Entendimento e análise descritiva dos dados
- Preparação de dados para mineração
- Mineração de regras de associação
- Agrupamento
- Regressão e Classificação

Bibliografia

- HAN, J., Kamber, M. and Pei, J.
  Data Mining - Concepts and Techniques. Morgan Kaufmann, 2011
- TAN, P. N., Steinbach, M. and Kumar, V.
  Introdução ao Data Mining - Mineração de Dados. Ciência Moderna, 2009.
- WITTEN, I. H., Frank, E. and Hall, M. A.
  Data Mining - Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2011.
- Artigos científicos de periódicos e eventos

Bibliografia



Bibliografia – Biblioteca Virtual

- Leandro Nunes de Castro e Daniel Gomes Ferrari
  Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações, 2016

- Clodis Boscarioli, Leandro Augusto da Silva e Sarajane Marques Peres
  Introdução à Mineração de Dados - Com Aplicações em R, 2016

- Mohammed J. Zaki e Wagner Meira Jr
  Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2020
  `https://dataminingbook.info`

Organização

- Ensino Remoto Emergencial (ERE)
  - Aulas assíncronas
  - Atendimento/discussões/apresentações de trabalho no horário da aula de quarta-feira
- Material
  - Videoaulas
  - Livros (biblioteca virtual)
  - Notas de aula (*Slides*)
  - Códigos

Avaliação

Avaliação	T1	T2	T3	T4	T5	T6	T7
Data	06/01	20/01	03/02	10/02	24/02	03/03	17/03
Valor	5	15	15	15	20	15	15
Sínc/Assínc	A	S?	A	A	S?	S	S

- Trabalhos
  - Apresentação do trabalho (com discussões/perguntas) no horário da aula nas quartas-feiras (S)
  - Texto explicando o que foi feito e discutindo os resultados – Google Classroom
  - Código fonte (não será avaliado) – Google Classroom
- Nota Final: soma das notas dos trabalhos
- Os trabalhos poderão ser feitos em duplas

Implementações e Experimentos Computacionais

- Google Colab
- Python
- Bibliotecas
  - numpy
  - scipy
  - matplotlib
  - pandas
  - scikit-learn
- Pode-se usar outras linguagens de programação, bibliotecas e ambientes de desenvolvimento

Introdução

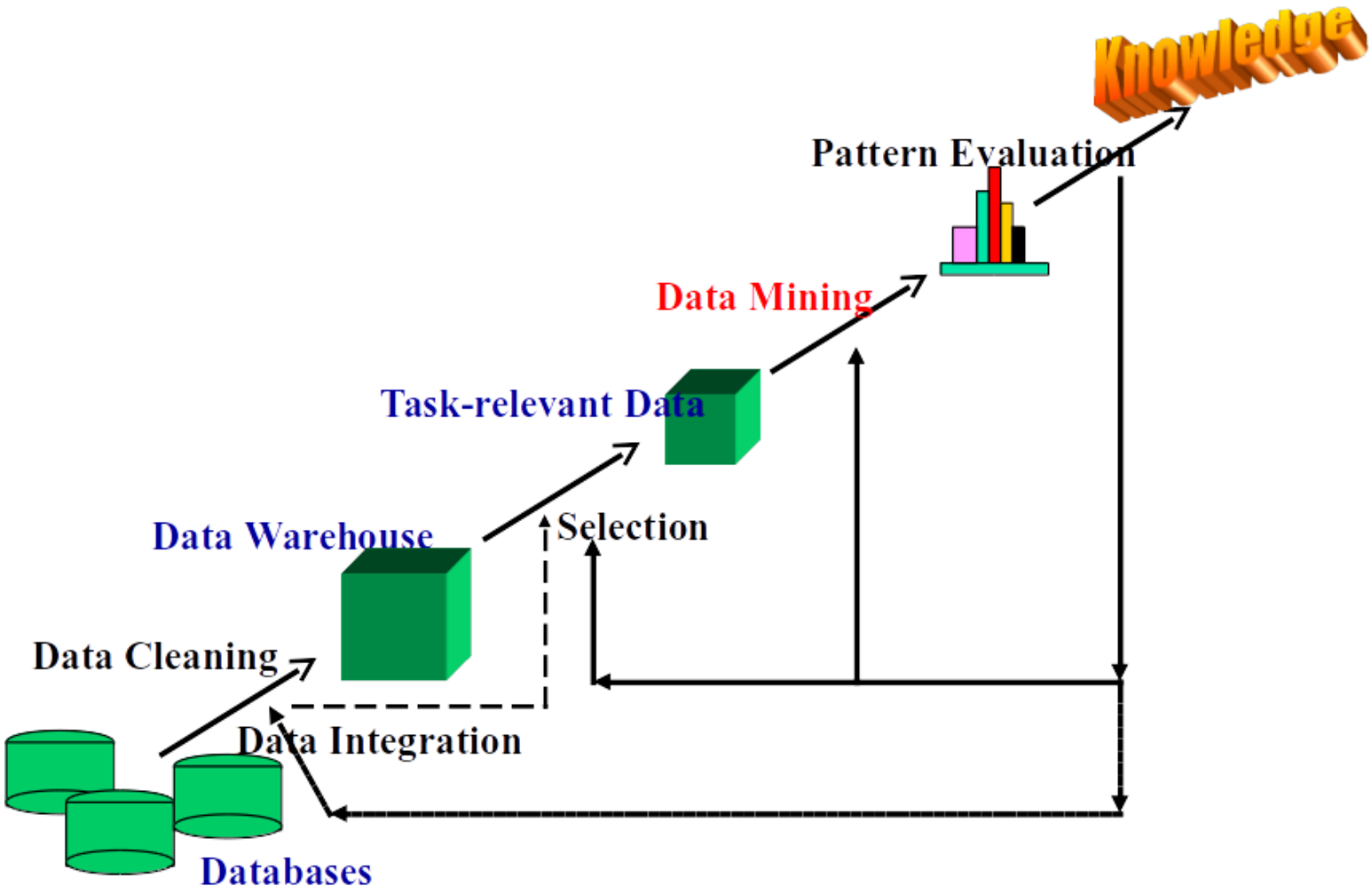
Por que Mineração de Dados?

- Houve um grande crescimento da quantidade de dados
- Muitos dados e busca por conhecimento
- A Mineração de Dados pode ser vista como uma automatização da análise de dados massivos
- A Mineração de Dados acompanha a evolução
  - da ciência
  - da tecnologia

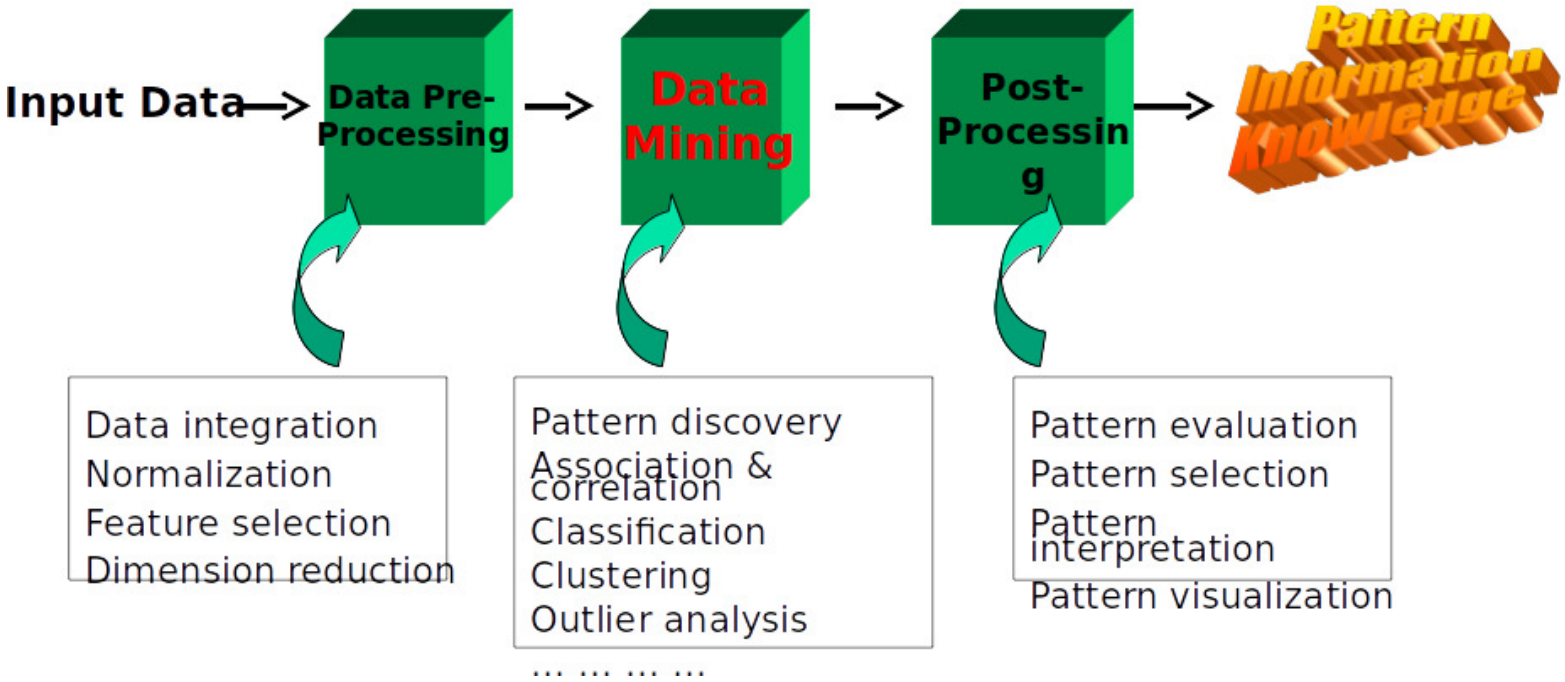
O que é Mineração de Dados?

- Mineração de Dados
- Descoberta de Conhecimento a partir de Dados
- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, business intelligence, *etc*
- Extração de padrões de interesse (não triviais, implícitos, desconhecimentos, potencialmente úteis) ou conhecimento a partir de uma grande quantidade de dados

Processo de Descoberta de Conhecimento



Processo – Aprendizado de Máquina



Exemplo: Análise de Dados Médicos

- Mineração de dados médicos ou de sistemas em saúde
- Pré-processamento dos dados
  - inclui extração de características e redução de dimensionalidade
- Processos de classificação e agrupamento
- Pós-processamento
  - apresentação dos resultados

Visão Multidimensional da Mineração de Dados

- Dados para serem minerados
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs and social, and information networks
- Conhecimento a ser minerado
  - Caracterização, discriminação, associação, classificação, agrupamento, tendência/derivação, análise de *outliers*, *etc*
  - Mineração de dados descritiva *vs.* preditiva

Visão Multidimensional da Mineração de Dados

- Técnicas utilizadas
  - Data warehouse (OLAP), machine learning, statistics, visualization, high-performance, etc
- Aplicações
  - Vendas, telecomunicações, bancos, análise de fraude, mineração de dados biológicos, análise de estoque, mineração de textos, mineração *web*, *etc*

Tipos de Dados

- Conjuntos e aplicações orientados a banco de dados
  - Relational database, data warehouse, transactional database
- Aplicações e conjuntos de dados avançados
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Associação e Correlação

- Padrões frequentes (ou *itemsets* frequentes)
  - Quais itens são frequentemente comprados juntos?
- Associação, correlação e causalidade
  - regra típica: frauda $\rightarrow$ cerveja
- Como minerar tais padrões e regras eficientemente em grandes quantidades de dados?

Classificação

- Aprendizado supervisionado
- Classificação
  - Modelos construídos usando dados de treinamento
  - Descreve e distingue as classes para predições futuras
- Exemplos de métodos
  - Árvore de Decisão, Naïve Bayes, Máquinas de Vetores de Suporte, Redes Neuronais Artificiais, Regressão Logística, Floresta Aleatória, ...
- Exemplos de aplicações
  - Detecção de fraude de cartão de crédito, direcionamento de *marketing*, identificação de doenças, ...

Agrupamento

- Aprendizado não supervisionado
- Os dados são agrupados para formar conjuntos sem haver supervisão
- Princípio: Maximar similaridade intra-grupo e minimizar a similaridade inter-grupos
- Exemplos de métodos
  - k-Médias, Aglomerativo, DBScan, ...
- Exemplos de aplicações
  - Detecção de anomalias, segurança, ...

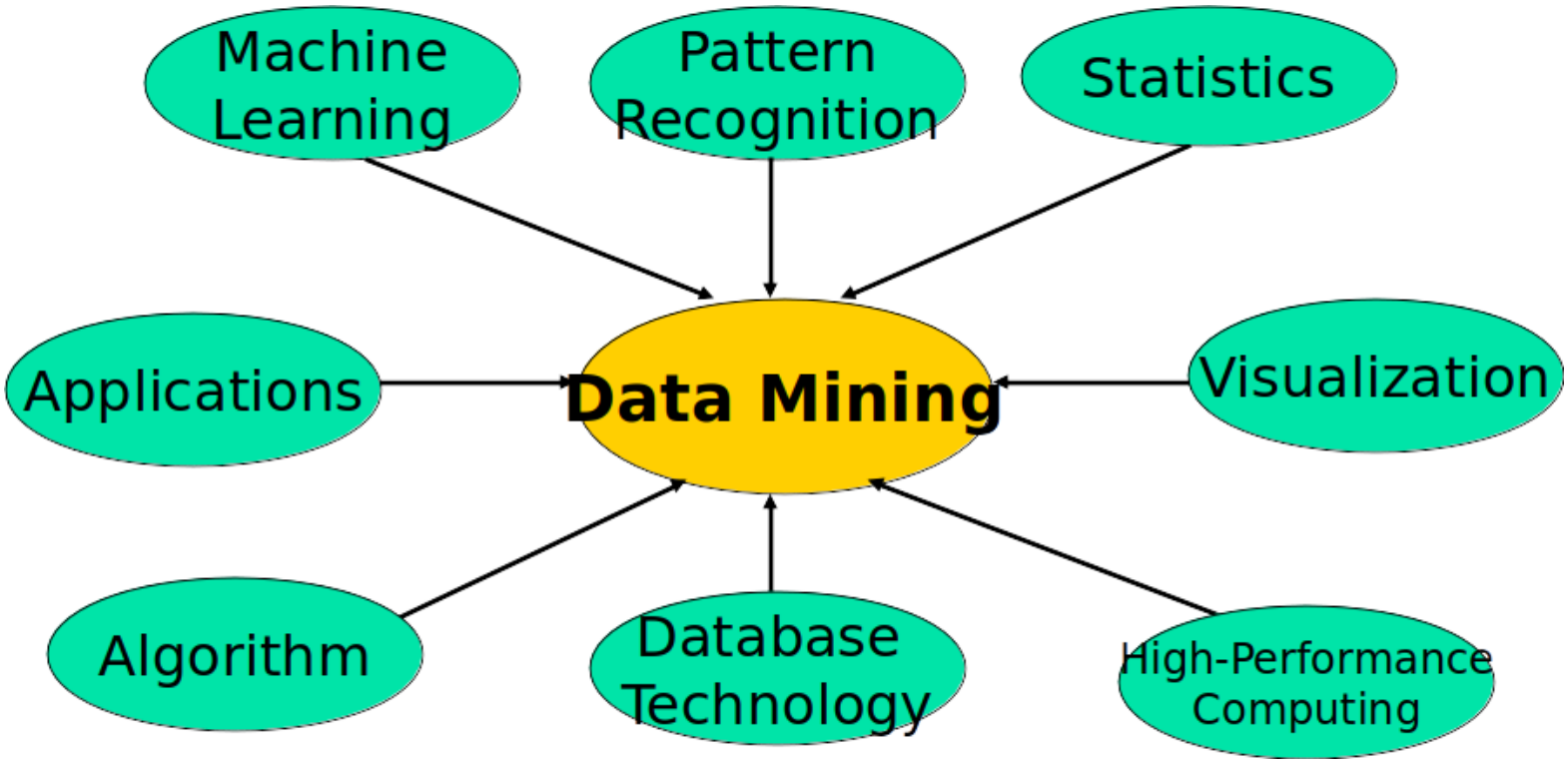
Análise de *Outlier*

- *Outlier*: Um dado que não está de acordo com o comportamento observado dos demais
- Ruído ou exceção?
  - O lixo de uma pessoa pode ser um tesouro para outra
- Métodos: agrupamento ou análise de regressão
- Útil em detecção de fraude e análise de eventos raros

Avaliação do Conhecimento

- Todos os conhecimentos minerados são interessantes?
  - Pode-se minerar uma quantidade MUITO grande de padrões e conhecimentos?
  - Alguns podem se limitar a certas dimensões do espaço (tempo, localização, etc)
  - Alguns podem não ser representativos, podem ser transientes, ...
- Avaliação dos dados minerados – minerar apenas conhecimento interessante?

Confluência de Múltiplas Disciplinas



Por que combina várias disciplinas?

- Grandes quantidades de dados
  - Algoritmos devem ser escaláveis para lidarem com grandes quantidades de dados
- Dados com muitas dimensões
  - *Micro-arrays* podem ter dezenas de milhares de dimensões
- Dados complexos
  - *Data streams*
  - Séries temporais e sequências de dados
  - Grafos, redes sociais e dados com múltiplas ligações
  - Bancos de dados heterogêneos e bases legadas
  - Bases de dados espaciais, multimídia, textos e *web*
  - Programas e simulações científicas
- Aplicações novas e sofisticadas

Conferências e Periódicos

- Data Mining and Knowledge Discovery (DAMI or DMKD)
- IEEE Trans. On Knowledge and Data Eng. (TKDE)
- KDD Explorations
- ACM Trans. on KDD
- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
- SIAM Data Mining Conf. (SDM)
- (IEEE) Int. Conf. on Data Mining (ICDM)
- European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)