



# MBA em Inteligência Artificial e Big Data

## – Curso 3: Administração de Dados Complexos em Larga Escala –

Caetano Traina Júnior

José Fernando Rodrigues Júnior

Robson Leonardo Ferreira Cordeiro

Grupo de Bases de Dados e Imagens – GBdl  
Instituto de Ciências Matemáticas e de Computação  
Universidade de São Paulo - São Carlos





# MBA em Inteligência Artificial e Big Data

## – Curso 3: Administração de Dados Complexos em Larga Escala –

### Técnicas avançadas para Preparação de Dados em SQL

Caetano Traina Júnior.

Grupo de Bases de Dados e Imagens – GBdl  
Instituto de Ciências Matemáticas e de Computação  
Universidade de São Paulo - São Carlos

Usando **Sistemas Gerenciadores de Bases de Dados** e **Repositórios de Dados** para a armazenagem, preparação e acesso a grandes volumes de dados – Conceitos gerais.



## Técnicas avançadas para Preparação de Dados em SQL – Introdução

Conceitos gerais em **Big Data**

Definição do ambiente de experimentação usado nesta parte do curso



# Roteiro

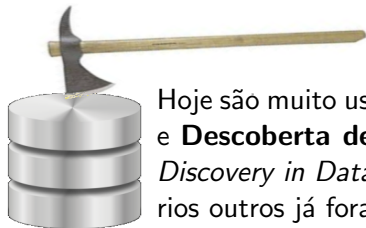


- 1 Conceitos básicos
- 2 Big Data
- 3 Ambiente de experimentação



# Descoberta de Conhecimento e Mineração de Dados

## Histórico



Hoje são muito usados os termos **Mineração de Dados** (*Data Mining*) e **Descoberta de Conhecimento em Bases de Dados** (*Knowledge Discovery in Databases – KDD*), mas na história da computação, vários outros já foram usados.



# Descoberta de Conhecimento e Mineração de Dados

## Histórico



- Hoje se usa muito os termos **Mineração de Dados** (*Data Mining*) e *Descoberta de Conhecimento em Bases de Dados* (*Knowledge Discovery in Databases*), mas vários outros foram usados no passado.
- Nos anos 1960, os estatísticos usavam “pescaria” ou “escavação” de dados (*Data Fishing* ou *Data Dredging*) para se referir – **depreciativamente** – às más práticas de analisar dados sem ter uma hipótese sobre o que se buscava.
- O termo *Database Mining: Mineração em Bases de Dados* ganhou força nos anos 1990 na comunidade de bases de dados. Mas ele foi registrado em nome da empresa Fico (*Fair Isaac Corporation*), e o termo acabou sendo substituído por *Mineração de Dados (data mining)*.
- Outros termos usados no passado incluem *Descoberta de informação*, *Extração de conhecimento*, *Information Harvesting*, etc.
- mas “*data mining*” se tornou popular no mercado e na imprensa leiga.



# Descoberta de Conhecimento e Mineração de Dados

## Histórico



- Gregory Piatetsky-Shapiro cunhou o termo **Descoberta de Conhecimento em bases de dados** (*Knowledge Discovery in Databases*) para criar o primeiro *workshop* sobre o tema em 1989, tornando esse termo popular nas comunidades de AI e aprendizado de máquina.
- Em 1996, Fayyad e colegas propuseram integrar uma terminologia em que **Mineração de Dados** especifica a fase mais elaborada da análise dos dados, como parte de um processo mais abrangente ao qual se associou o termo **Descoberta de Conhecimento em Bases de Dados**, embora frequentemente ambos continuem sendo usados indistintamente.



# Descoberta de Conhecimento e Mineração de Dados

## Histórico



- A Mineração de Dados é fundamentada em técnicas de Inteligência Artificial e Análise Estatística.
- Com a coleta e armazenagem de dados propiciadas pelas tecnologias de Bases de Dados, a integração das três tecnologias passaram a dar resultados excepcionais, e o interesse de sua aplicação nas mais diversas áreas de atividades explodiu.
- A integração das três áreas de conhecimento passou a ser chamada de Ciências de Dados e de Engenharia de dados:







### Ciências de Dados

É o processo de criação e treinamento de modelos preditivos, usando dados preparados em formatos adequados para a análise, e sua integração com as tarefas de especificação, interpretação e entendimento dos resultados pelos gerentes e executivos de um processo produtivo.

### Engenharia de dados

É o conjunto de tecnologias para a construção e manutenção dos sistemas que permitem aos cientistas de dados acessar e analisar os dados, executando as atividades de coleta, armazenagem, recuperação, limpeza e transformação de dados, dando apoio à construção de *pipelines* de dados e gerenciamento da procedência e reprodutibilidade de todas as atividades sobre os dados e as informações extraídas.

# Descoberta de Conhecimento e Mineração de Dados

## Motivação



### MBA em Inteligência Artificial e Big Data

#### – Curso 3: Administração de Dados Complexos em Larga Escala –

Visa abordar atividades de **Engenharia de Dados** na armazenagem, recuperação e preparo de dados, ilustrando técnicas que enfatizam a **eficiência** dos processos para tratar **grandes volumes de dados**.

### Parte 1

Aprender técnicas avançadas para  
**Preparação de Dados em SQL**



### Compartilhamento de dados

O grande avanço científico em tecnologias de informação se deve em grande parte ao compartilhamento de experiências, software e **dados**.

- Promove transparência
- encoraja a colaboração
- Acelera a pesquisa
- Melhora os processos de tomada de decisão

- Em áreas como a **saúde pública**, e em particular em emergências como surtos de doenças infecciosas tal como é a recente situação de pandemia, o compartilhamento de dados tem sido fundamental;
- Como **profissionais de computação**, nos beneficiamos desse avanço.

Portanto, temos a responsabilidade ética de **promover o compartilhamento** dos dados **que temos** (dentro do possível), adotando iniciativas como **open data e fair data**.

# Descoberta de Conhecimento e Mineração de Dados

## Motivação



- O que é **FAIR data**?
- É um princípio de compartilhamento de dados promovido por um consórcio de cientistas de todo o mundo: <https://www.force11.org/fairprinciples>  
(**FAIR** Findable, **A**ccessible, **I**nteroperable, and **R**eusable)

**Localizável** - dados e material complementar deve ter metadados suficientemente ricos (*dspace*), bem como um identificador único e persistente, como DOI.

**Acessível** - dados metadados devem ser compreensíveis para humanos e máquinas, e ficar disponíveis em um repositório confiável.

**Interoperável** - os metadados devem usar uma linguagem formal, acessível, compartilhada e amplamente utilizável em diversas plataformas e eco-sistemas.

**Reutilizável** - os dados e coleções devem ter uma licença de uso clara e fornecer informações precisas sobre sua procedência e uso responsável.





### Dados abertos × Dados Fair

- **Dados FAIR** e **Dados Abertos** são conceitos distintos:
- Por exemplo, em medicina e em saúde pública, os dados dos pacientes e as informações de identificação pessoal (*Personally Identifiable Information* – PII), podem ser **FAIR** mas não **abertos**, para garantir a proteção, privacidade e confidencialidade das pessoas;
- Nesse caso, os metadados e informações agregadas podem ser disponibilizadas publicamente **FAIR**, mas os dados brutos originais não podem ser **abertos**.



# Descoberta de Conhecimento em Bases de Dados

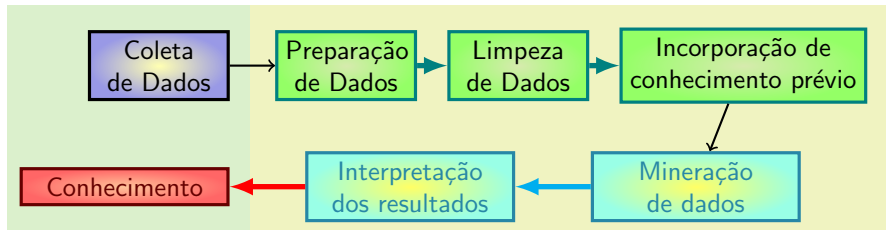
## Terminologia

Knowledge discovery in databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Usama Fayyad-1996].

“Descoberta de conhecimento em bases de dados é o processo não trivial de identificar padrões nos dados, que sejam válidos, inéditos, potencialmente úteis e em essência compreensíveis.”

A partir dos dados disponíveis em um *Data Warehouse/Data lake*,

o processo de KDD é realizado em várias fases de maneira iterativa.



# Descoberta de Conhecimento e Mineração de Dados

## Data warehousing



### Coleta de dados

- Os dados devem ser previamente coletados e disponibilizados de maneira estável, em geral de maneira acumulativa.
- Os dados coletados em *Data warehouses* – e mais recentemente, em *Data lakes* – se destinam à análise, ou seja, os dados são apenas acumulados, mantidos estáveis e não mais sofrem atualizações, visando a repetitividade das análises.

👉 *Data warehouses/lakes* correspondem a prover um empreendimento com **memória**.

👉 *Data mining* corresponde a prover um empreendimento com **inteligência**.

# Descoberta de Conhecimento em Bases de Dados



## *Data warehouses* × *Data lakes*

- Tanto em **data warehouses** quanto em **data lakes**, os dados em geral se destinam à análise, e portanto não são atualizados, apenas acumulados.
- Em um **data warehouse**, os dados são coletados e disponibilizados de maneira **estruturada e estável**, em geral **acumulando os dados transacionais** do empreendimento, registrando a história da evolução dos dados.
- Em um **data lake**, armazenam-se também **fragmentos de dados** obtidos durante os processos de coleta e processamento de dados, em geral de maneira **não estruturada**, **sem planejamento**, incluindo dados brutos de sensores, resultados intermediários, *logs*, etc.

👉 *Data warehouses* tendem a ser usados por ferramentas baseadas em regras pré-definidas, em especial em tarefas de **Business Intelligence**.

👉 *Data lakes* tendem a ser usados também por ferramentas de exploração de conhecimento, tal como em tarefas de **Mineração de Dados**.





### O processo de Descoberta de Conhecimento e Mineração de Dados – KDD –

- A **Descoberta de Conhecimento e Mineração de Dados** envolve diversas áreas do conhecimento:
  - **Bases de dados;**
  - **Inteligência Artificial**/Aprendizado de máquina+Reconhecimento de padrões;
  - **Estatística;**
  - **Teoria da Informação;**
  - **Visualização** de dados e de Informação;
  - **Computação de alto desempenho;**
  - **... e as regras do negócio da aplicação!**



# Descoberta de Conhecimento e Mineração de Dados

## O Papel da área de Bases de Dados



Dentre as várias áreas que contribuem com a **Mineração de Dados**, a área de **Bases de Dados** contribui principalmente com:

- **Escalabilidade dos processos**
  - Quanto ao número de instâncias (cardinalidade);
  - Quanto ao número de atributos envolvidos (dimensionalidade);
  - Quanto à quantidade de informação representada (domínio de dados).
- **Automatização dos processos**
  - reduzir a necessidade de intervenção do usuário;
  - Administrar a reprodutibilidade dos experimentos.
- E permitir **trabalhar com dados em volumes muito grandes e muito heterogêneos**
- **Big Data**



# Descoberta de Conhecimento e Mineração de Dados

## O Papel da área de Bases de Dados



A escalabilidade é procurada em quatro abordagens distintas:

- Desenvolvendo algoritmos mais eficientes, com menor complexidade computacional — usando técnicas de indexação e pré-computação dos dados que serão intensamente consultados;
- Usando técnicas de paralelização e distribuição de dados;
- Usando técnicas de particionamento de dados;
- Usando técnicas de representação relacional de dados, incluindo a normalização multi-relacional.

👉 Os dados são constituídos de grandes “conjuntos”, portanto a representação é relacional. Não significa que sua armazenagem e recuperação esteja subordinada ao uso de SQL:

**Relacional** × **NoSQL** × **NewSQL**!



- Dados que não forem armazenados, estarão perdidos para sempre  
👉 Experimentos muito caros, Dados meteorológicos, Atividades socioeconômicas, Agronomia...
- Técnicas de análise na dimensão temporal têm maior probabilidade de acerto se houver dados com uma longa história...
- “Não sei que dados vou precisar amanhã, então vou guardar tudo o que tenho hoje...”

Bases de Dados Transacionais (OLTP) × Bases de Dados Analíticas (OLAP)    👉 *Data warehouses*

Bases de Dados não-estruturados e semi-estruturados    👉 *Data lakes*

# Big Data



O termo **Big Data** foi inicialmente cunhado por um grupo de consultoria em tecnologia da informação americano (Gartner Inc.) em 2001, para alavancar a prospecção de oportunidades de negócios e desenvolvimento de tecnologia:

## *Gartner definition (2012)*

*"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."*

- Software,
- Infra-estrutura de data-centers,
- Telecomunicações.



# Big Data

Um empreendimento é considerado lidar com **Big Data** quando existem desafios em ao menos três “**V-dimensões**” (the **Big Vs**):





Muitos gostam de adicionar outras **V-dimensões**:

**Veracidade** - Indica quanto os dados coletados de fato refletem a realidade;

**Valor** - Indica quanto os dados são úteis para o empreendimento.

Estas dimensões se referem à **confiabilidade** do processo de coleta dos dados, à capacidade da empresa de se beneficiar dos dados.

👉 Elas estão mais ligadas ao contexto onde ocorrem e menos ao que os dados de fato **são**.

- Bases de dados e ferramentas de análise não são diretamente afetadas por elas.





Existem diversas abordagens genéricas para tratar *Big Data*, e alguns termos são muito usados hoje:

- Processamento maciçamente paralelo (MPP),
- *Crowdsourcing*,
- Simulação,
- Redução de Dados







Mas... **Quão BIG é BIG?**


quer dizer: **Quão grande os dados precisam ser para se falar em Big Data?**

★ Quando qualquer das V-dimensões exceder a capacidade do sistema para processar os dados com um **uso aceitável dos recursos disponíveis**:


- memória de trabalho,
  - memória de armazenagem,
  - tráfego em rede, ...
  - mas, principalmente: **tempo de processamento** !
- **Volumes** relativamente pequenos (poucos GBytes ou até mesmo MBytes) podem ser **big data**:⇒ se precisarem
- de muito **tempo** de processamento ou
  - ser armazenados muito rapidamente, **(velocidade)** ou
  - interpretar muitos formatos diferentes **(variedade)** ...





O problema é manter a **velocidade** para se obter a resposta mesmo quando o **volume** dos dados aumenta,  e um grande volume pode ocasionar também o aumento de sua **variedade**, **velocidade de acesso** individual, etc.

Portanto, embora muitos **Vs** possam ser considerados, o problema está em manter a **escalabilidade** do sistema em relação ao **volume dos recursos** necessários.

 O **volume** dos dados aumenta em **taxas superlineares**, e a maioria dos recurso (especialmente **hardware**) aumentam em **taxas sublineares**.

As únicas soluções reais envolvem necessariamente o desenvolvimento de **tecnologias escaláveis** baseadas em **lógica** para o tratamento de dados:

★ software ★,

- e devem ser específicas para cada problema.



## A solução depende:

tanto de quais recursos são necessários e quais podem ser alocados,  
quanto das características dos próprios dados e do uso pretendido.

- Para Alta taxa de leitura (ou escrita) em memória secundária:  
é necessário **particionar** o acesso 🖱️ arquiteturas Map-Reduce;
- Para Alta disponibilidade: se é necessário ter o sistema sempre disponível:  
é necessário **distribuir** os dados 🖱️ arquiteturas baseadas em *sharding*;
- Mas se a **complexidade computacional** é elevada:  
é necessário aumentar a **eficiência** 🖱️ algoritmos eficientes!
  - O problema é recuperar partes dos dados 🖱️ **indexação**
  - O problema requer ler todos os dados 🖱️ **reduzir, particionar**
  - O problema requer cruzar dados 🖱️ **correlações, Teoria de fractais**





- Portanto, soluções baseadas em aumentar o **hardware** disponível podem ser usadas de maneira temporária.
- Mas a solução efetiva para o problema da **escalabilidade** do processamento de **Big Data** depende essencialmente de **software**, possivelmente apoiado em **hardware especializado**
- As principais táticas de **software** incluem:
  - Redução de dados
  - Indexação e particionamento
  - Estruturas de dados e pré-processamento
- Os principais módulos de **hardware especializado** incluem:
  - Unidades de processamento vetorial (GPUs)
  - Unidades de processamento tensorial (TPUs)
  - ... computação quântica?

E **software especializado**





- Nesta primeira parte do curso, será utilizado o

## Sistema de Gerenciamento de Bases de Dados Relacional



para a execução dos exemplos.

- Serão usadas três bases de dados:
  - **Alunos** — sintética — uma base minúscula, para ilustrar conceitos gerais: 15 alunos, 40 matrículas, etc.
  - **Alunos80K** — sintética — uma base mediana, simula a base de alunos da USP: 80.000 alunos, 720.000 matrículas, etc.
  - **FapespCovid-19** — real — uma base “grandinha”, contém 3 tabelas: 47.971 pacientes, 6.855.217 resultados de exames, 260.682 desfechos de internações

Essa base está disponível no repositório COVID-19 Data Sharing/BR - USP, em <https://repositoriodatasharingfapesp.uspdigital.usp.br/>.

Para os exemplos, foram carregados apenas os dados dos hospitais que disponibilizam desfechos: Hospital Sírio-Libanês e Hospital Beneficência Portuguesa em SP.

MBA em Inteligência Artificial e Big Data  
– Curso 3: Administração de Dados Complexos em Larga Escala –  
Técnicas avançadas para Preparação de Dados em SQL

Caetano Traina Júnior.

Grupo de Bases de Dados e Imagens – GBdI  
Instituto de Ciências Matemáticas e de Computação  
Universidade de São Paulo - São Carlos

Conceitos gerais  
**FIM**

