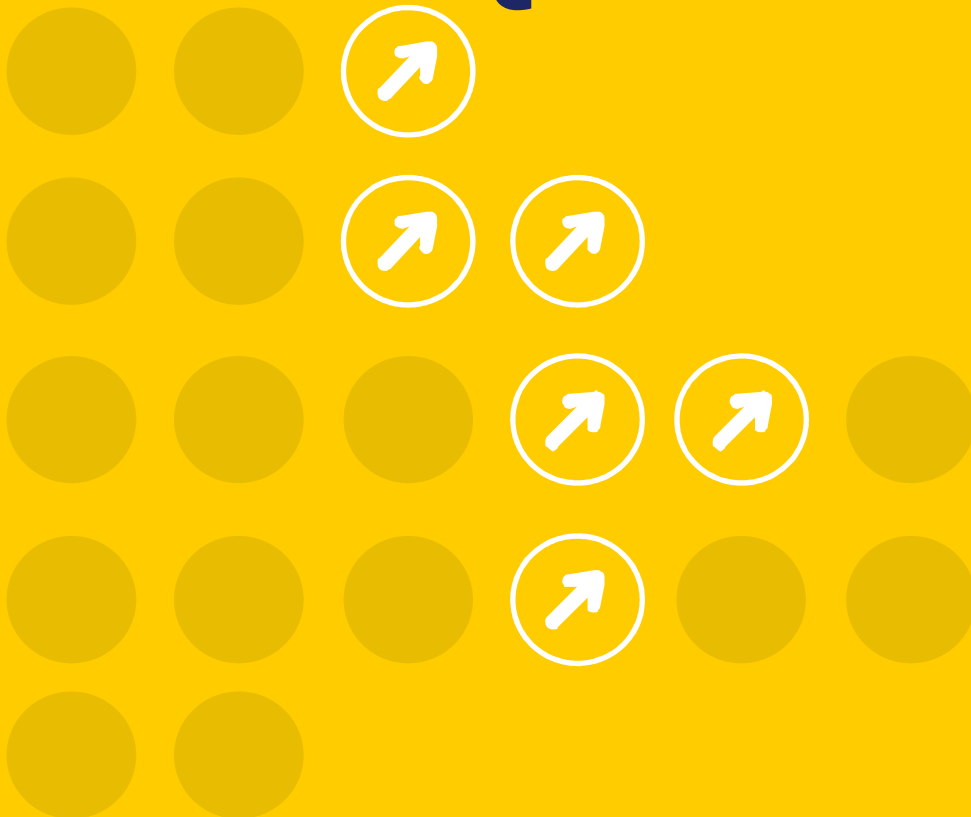


Banco de Dados, Integração e Qualidade de Dados



Ceça Moraes

cecafac@gmail.com



Sobre a professora

Ceça Moraes

↗ Doutora em Computação (UFPE)

Áreas de atuação

↗ Desenvolvimento de Software e Banco de Dados

Experiência (~25 anos):

↗ Várias empresas

↗ 2010 até hoje -> UFRPE

↗ Pesquisadora ISI-TICS (Senai)

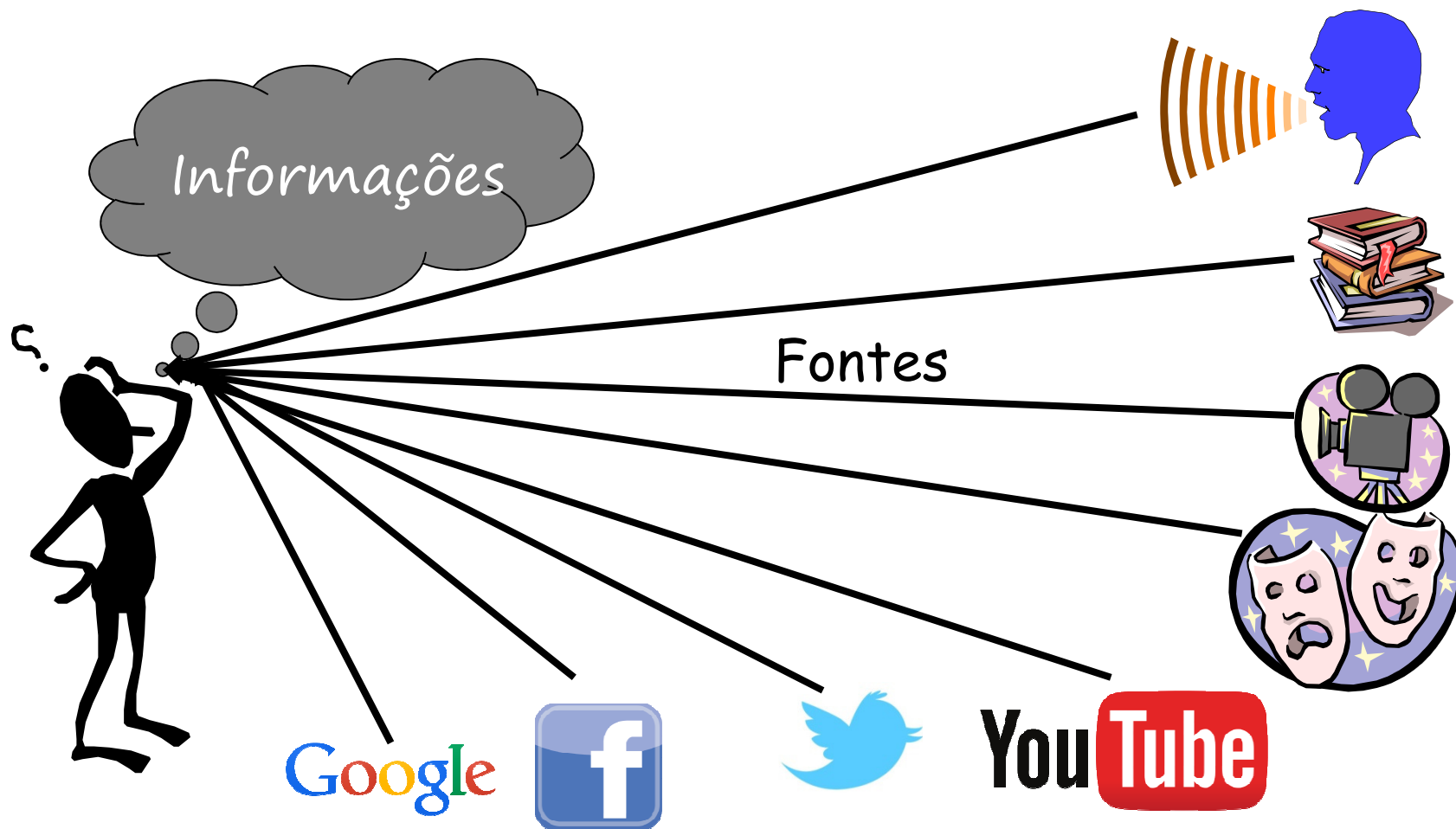


Conteúdo

- Apresentação
- Banco de Dados
 - ↗ Introdução
 - ↗ Conceitos Básicos
 - ↗ Modelos
 - ↗ Aplicações
 - ↗ Fundamentos
- Integração de Dados
- Qualidade da Informação

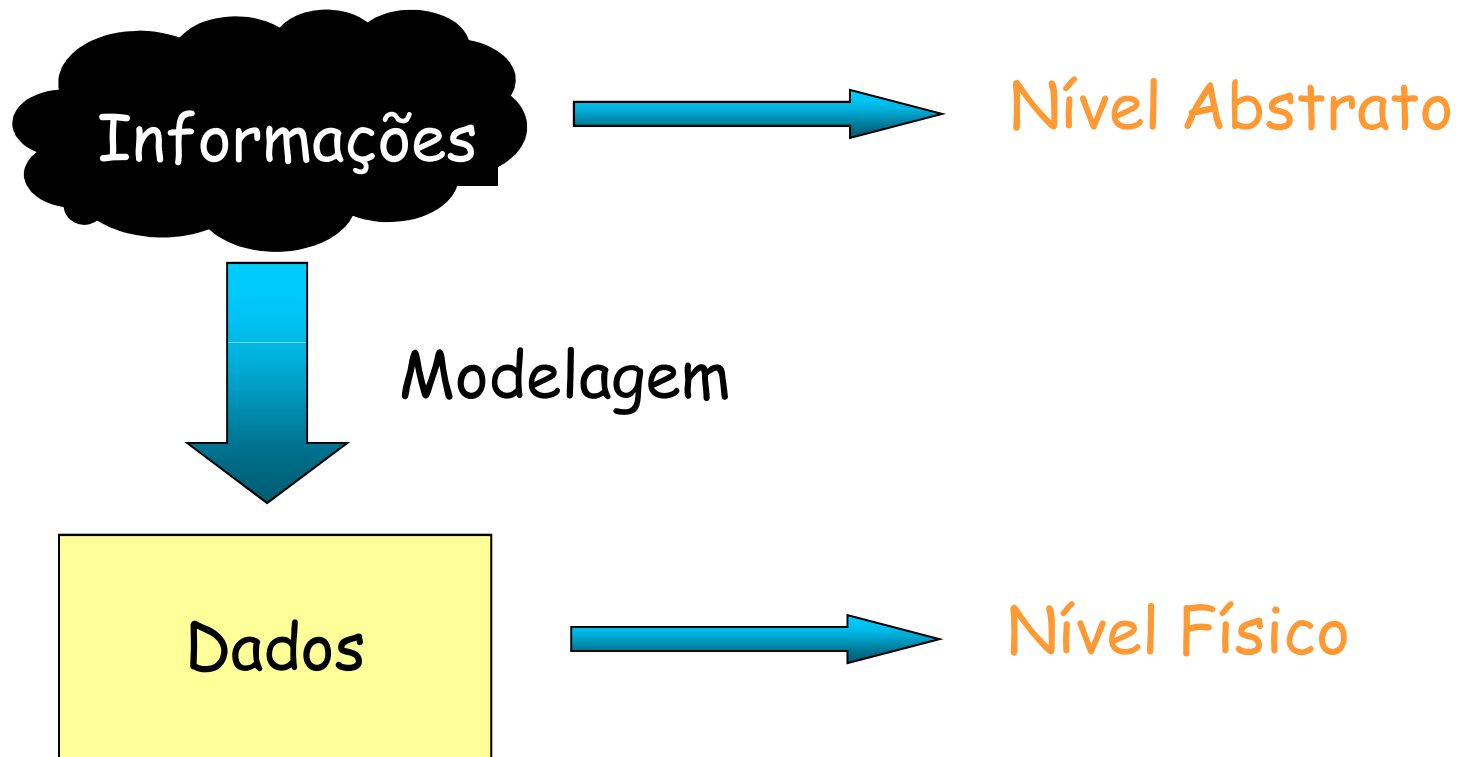


Estruturando Informações...





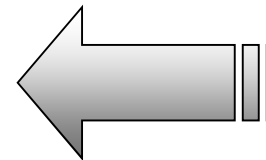
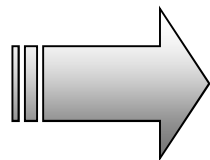
Estruturando informações...





Estruturando informações...

Dados



*Instruções
de
programas*



Processamento

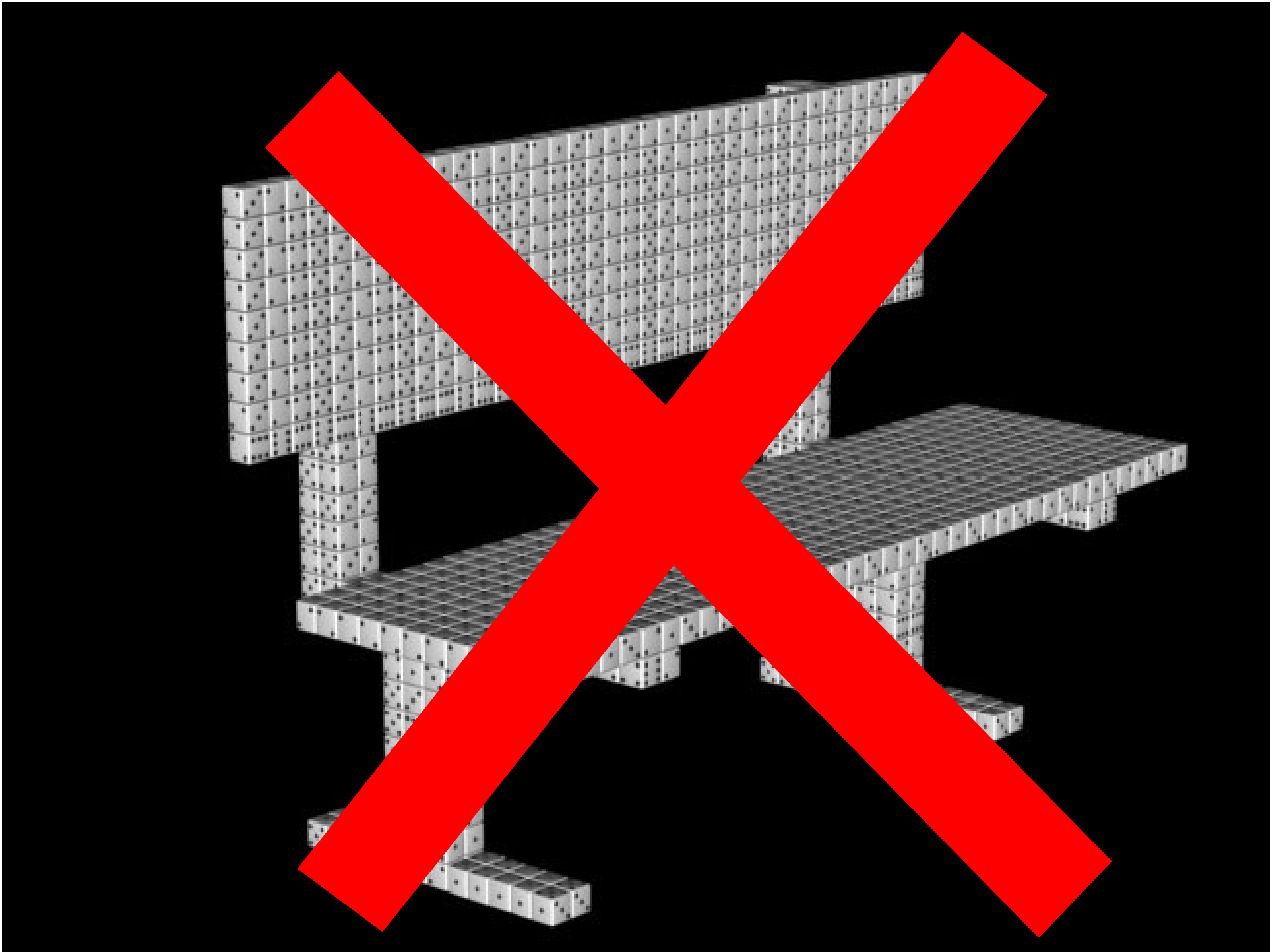
↗ Coletar, organizar, armazenar e recuperar informações de forma eficiente



Como fazer ????

Banco de Dados

- Conceitos -





Banco de Dados (BD)

- ❏ Conjunto de dados **inter-relacionados** para atender **necessidades** de um conjunto de usuários e/ou aplicações
- ❏ Conjunto de programas de **acesso** aos dados
 - ↪ *Sistema de Gerenciamento de Banco de Dados (SGBD)*



Antes...

- ❏ No início, aplicações usavam sistemas de **arquivos**
- ↪ Dados de diferentes aplicações **não integrados**
- ↪ Dados projetados para atender **a uma aplicação específica**



Antes...

❏ Problemas:

↗ **Redundância** e inconsistência

↗ Dificuldades no **acesso** aos dados

↗ Problemas de **compartilhamento**

↗ **Segurança**



Processamento sem BD

Exemplo: empresa hipotética implementa sistemas para

↗ Vendas

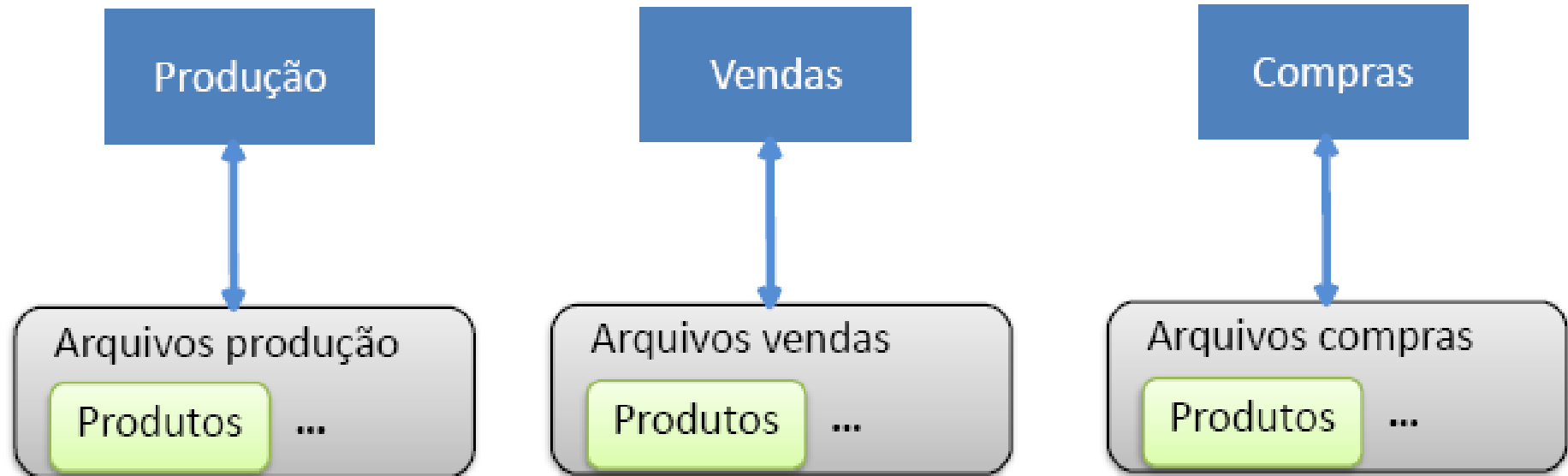
↗ Produção

↗ Compras

Exemplo: Onde ficam os dados de produtos?



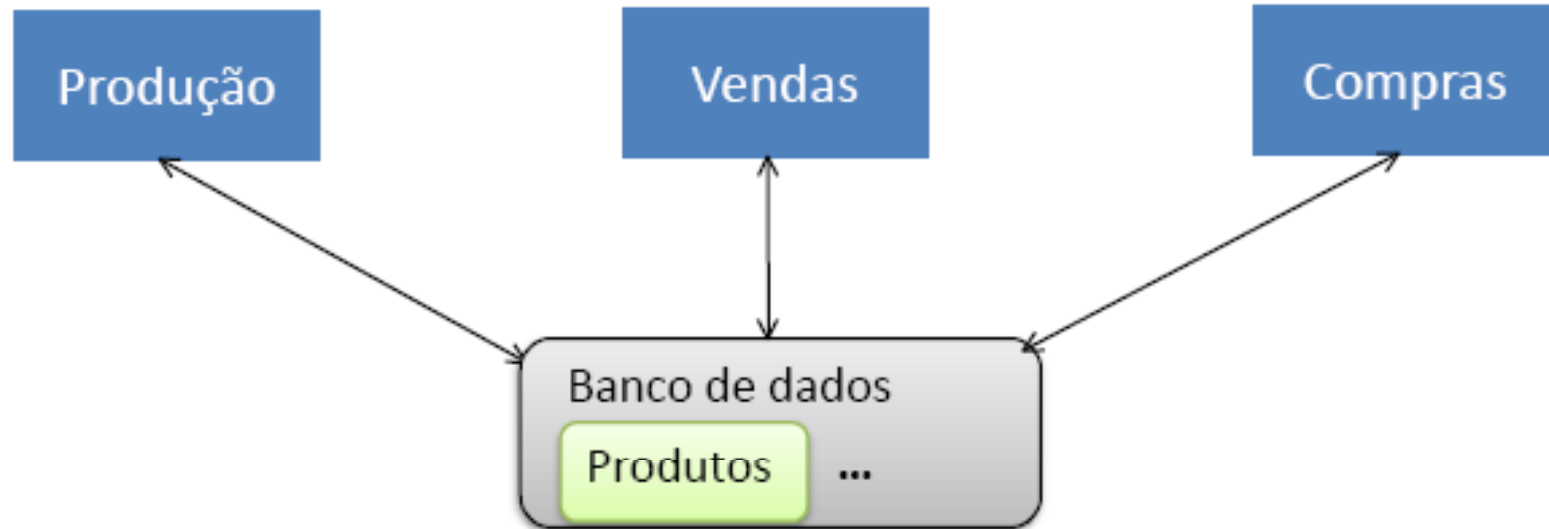
Processamento sem BD



❏ Problema de Redundância de dados



Processamento com BD



- 🖥️ Dados armazenados uma única vez:
 - ↪ Eventual redundância controlada pelo SGBD e invisível às aplicações



SGBD – o que é isso ?

- ❏ SGBD = **Sistema Gerenciador de Banco de Dados** (Database Management System)
- ❏ Um software sofisticado para:
 - ↪ **Armazenar** dados de forma eficiente
 - ↪ **Consultar** dados obtendo respostas rápidas
 - ↪ **Modificar** dados concorrentemente
- ❏ Exemplos: DB2 – IBM; Oracle Server – Oracle; SQL Server – Microsoft; MySQL; PostgreSQL



Características de um SGBD

- ☐ Integridade / Consistência
- ☐ **Restrições**
- ☐ Segurança / **Privacidade**
- ☐ Restauração
- ☐ **Reorganização**
- ☐ **Eficiência**



Modelos de SGBDs

Modelo

➤ Descrição formal dos dados armazenados em um banco de dados

Hierárquicos

Em **Rede**

Relacional

OO/OR



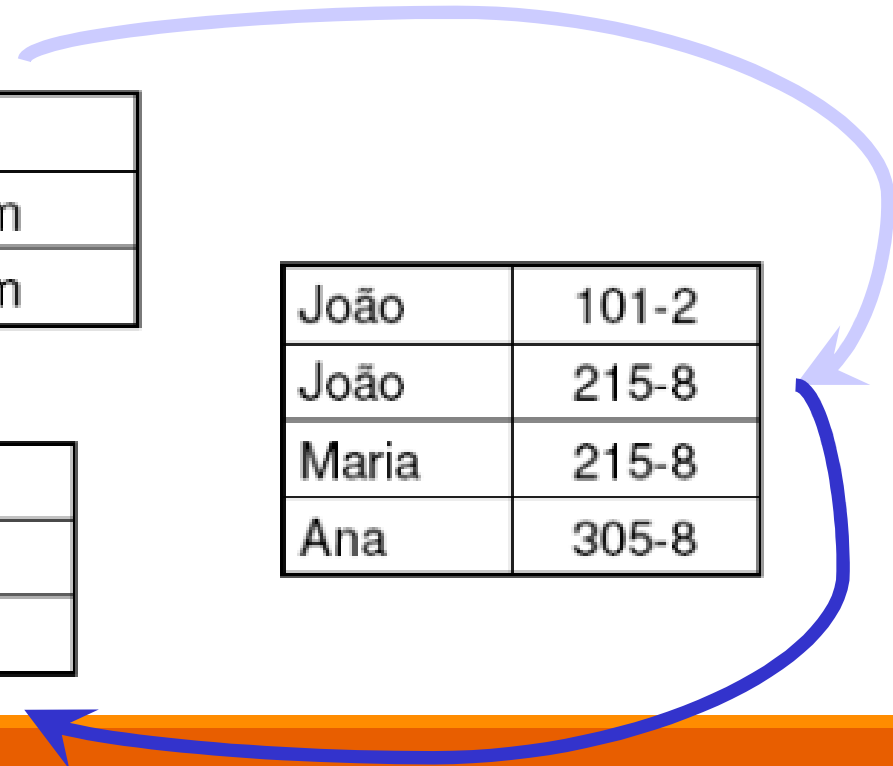
Modelo Relacional

- ❏ Dados e **relacionamentos** representados por **tabelas**
- ❏ **Tabelas possuem múltiplas colunas e linhas**

João	Centro	Natal
Maria	Sul	Parnamirim
Ana	Norte	Parnamirim

101-2	500
215-8	700
305-8	400

João	101-2
João	215-8
Maria	215-8
Ana	305-8



DESTINATARIO	
id	INT
nome_razasocial	VARCHAR(255)
cpf	VARCHAR(11)
endereco	VARCHAR(100)
bairro	VARCHAR(60)
cep	VARCHAR(8)
municipio	VARCHAR(255)
telefone	VARCHAR(11)
uf	CHAR(2)
pais	VARCHAR(45)
indicador_ie	VARCHAR(45)
inscricao_estadual	VARCHAR(45)
inscricao_suframa	VARCHAR(45)
im	VARCHAR(45)
email	VARCHAR(100)

NFE	
id	INT
chave	BIGINT
modelo	INT(2)
serie	INT(1)
numero	INT(9)
data_emissao	DATE
data_hora_saida_entrada	TIMESTAMP
valor_total	DECIMAL
emitente_id	INT
destinatario_id	INT
transporte_id	INT

PRODUTO	
id	INT
nome	INT
descricao	VARCHAR(255)
codigo_produto	INT

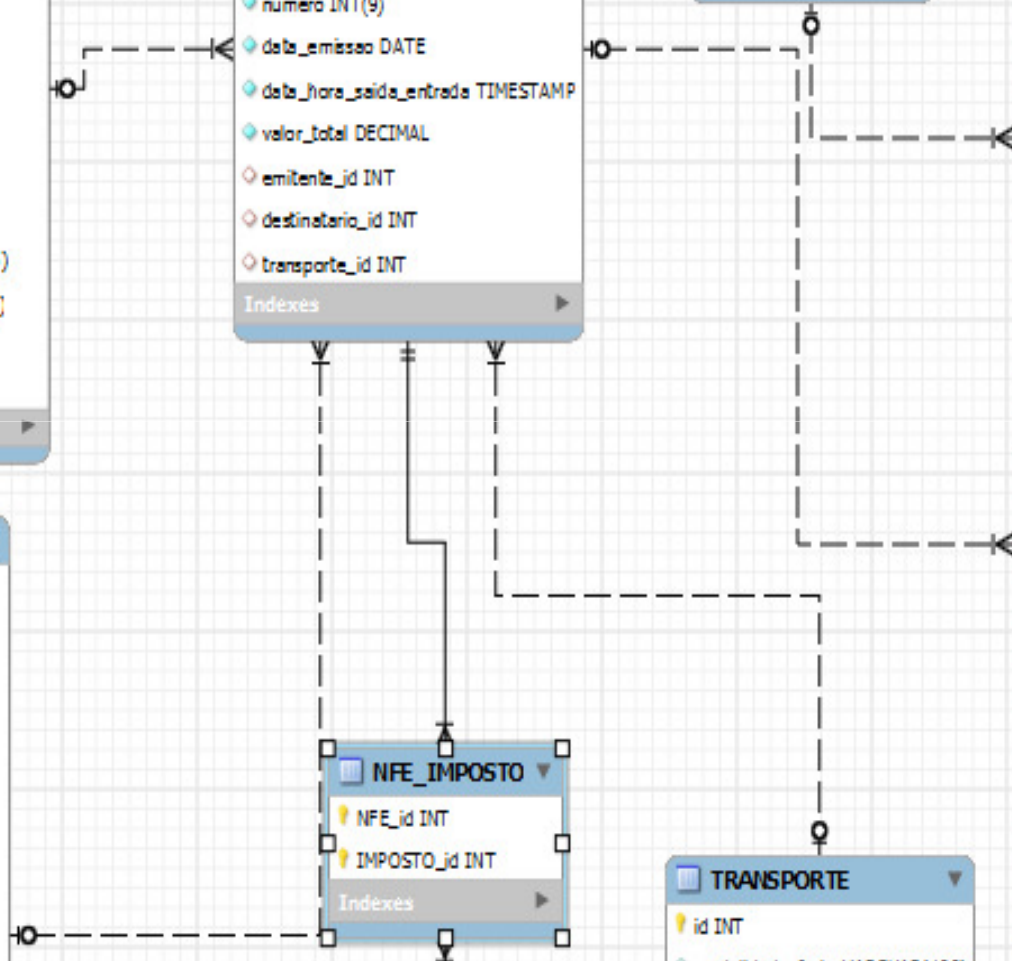
ITEM_NF	
id	INT
chave	BIGINT
quantidade	INT
valor	DECIMAL
codigo_nam	INT
codigo_ex_tipo	INT
cfop	INT
outras_despesas_acessorias	VARCHAR(255)
valor_desconto	DECIMAL
valor_total_frete	DECIMAL
valor_seguro	DECIMAL
indicador_composicao_valor_total_nfe	VARCHAR(255)
codigo_ean	INT
unidade_comercial	CHAR(3)
quantidade_comercial	DECIMAL
codigo_ean_tributavel	VARCHAR(14)
unidade_tributavel	CHAR(3)
quantidade_tributavel	DECIMAL
valor_unitario_comercializacao	DECIMAL
valor_unitario_tributacao	DECIMAL
numero_pedido_compra	INT
item_pedido_compra	INT
valor_aproximado_tributos	DECIMAL
numero_fci	INT
PRODUTO_id	INT
NFE_id	INT

EMITENTE	
id	INT
razasocial	VARCHAR(45)
nome_fantasia	VARCHAR(45)
cnpj	VARCHAR(14)
endereco	VARCHAR(100)
bairro	VARCHAR(80)
cep	VARCHAR(9)
municipio	VARCHAR(255)
telefone	VARCHAR(11)
uf	CHAR(2)
pais	VARCHAR(45)
inscricao_estadual	VARCHAR(12)
inscricao_estadual_subs_tributaria	VARCHAR(30)

NFE_IMPOSTO	
NFE_id	INT
IMPOSTO_id	INT

TRANSPORTE	
id	INT
modalidade_frete	VARCHAR(100)

IMPOSTO	
id	INT
base_calculo	DECIMAL





Exemplo BD Relacional

Passageiros

Nome	Telefone
Jose Campos	322-9999
Maria Silva	222-3333
Gabriel Silva	222-3333
Cláudio Silva	222-3333

Reservas

Nome Vôo	Passageiro
Air France 147	José Campos
Air France 147	Maria Silva
Air France 147	Gabriel Silva
Varig 224	Gabriel Silva
Varig 224	Cláudio Silva

Vôos

Vôo	Tarifa	Ass	Cia
Air France 147	1000,00	250	Air France
Air France 455	750,00	500	Air France
Varig 224	500,00	150	Varig

Cia

Nome	Faturam.	Presid.	Sede
Varig	20000K	Pampa	Porto Alegre
Air France	10000K	Etoile	Paris

Colunas

↪ Campos ou Atributos

Linhas

↪ Registros (dados)



SQL (Structured Query Language)

- ❏ **Linguagem** padrão de interação com banco de dados relacionais
 - ↗ Reconhecida por quase todos os produtos
 - ↗ Padrão ANSI (American National Standard Institute)
- ❏ Comandos para:
 - ↗ **Inserção, modificação, remoção** de **dados, consultas**
 - ↗ **Manipulação de estruturas**



Várias Aplicações e Tipos de SGBD

- Business Intelligence
- Bancos de Dados Geográficos
- Bancos de Dados Móveis
- Bancos de Dados Biológicos
- Cloud Computing
- BIG DATA!!!!!!!!!!!!!!**



Dados, dados e mais dados...

Ponto de Vista Comercial

- Quantidades gigantescas de dados são coletados e armazenados em empresas, corporações
 - Dados de comércio eletrônico
 - Dados de navegação na internet
 - Dados de compras de clientes em grandes lojas de departamentos, supermercados
 - Dados de transações bancárias ou de cartão de crédito

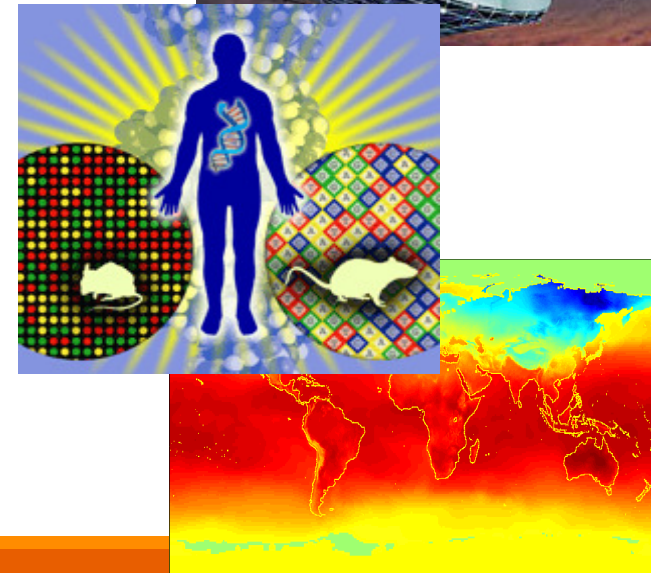
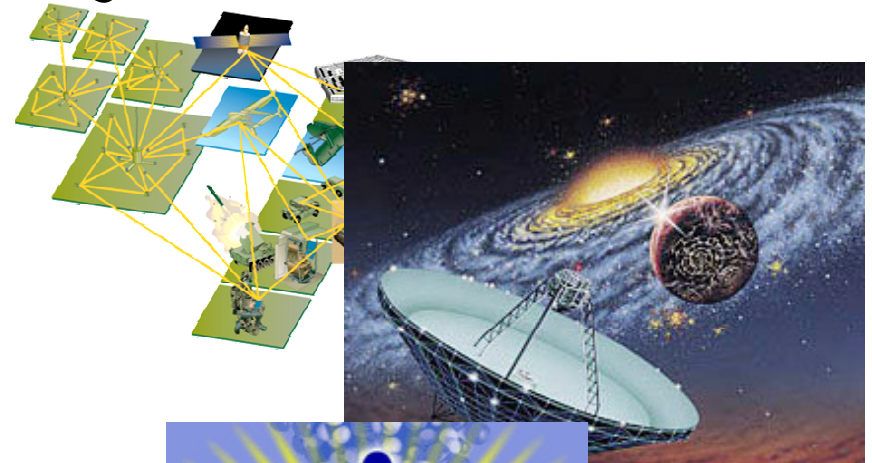




Dados, dados e mais dados...

- Dados coletados e armazenados a velocidades enormes (GB/hora)
 - Sensores remotos em satélites
 - Telescópios
 - Microarrays gerando dados de expressões de genes
 - Simulações científicas gerando terabytes de dados

*Ponto de Vista Científico:
Medicina, Biologia,
Engenharia*



Tá ok, mas como eu pego meu milhãozinho?

Tecnologias mudam rápido,
mas os fundamentos
permanecem

- Isto não quer dizer que você não vai aprender novas tecnologias
- Quer dizer que você **DEVE** aprender os fundamentos computacionais

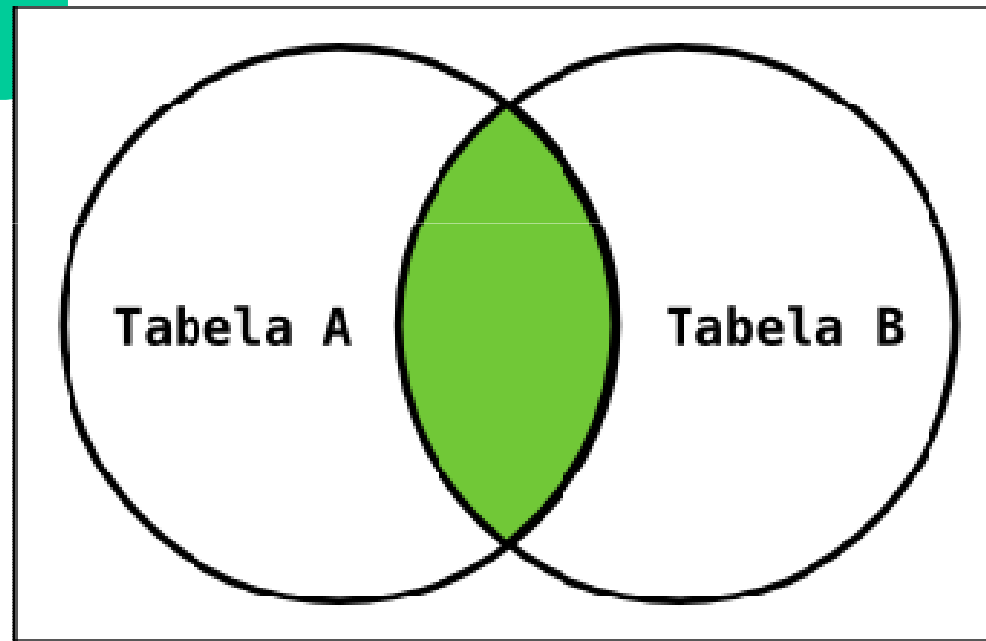
Use as tecnologias para estudar
os fundamentos!



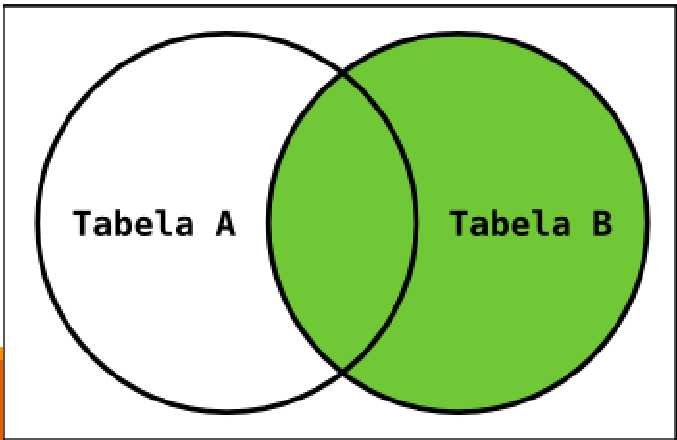
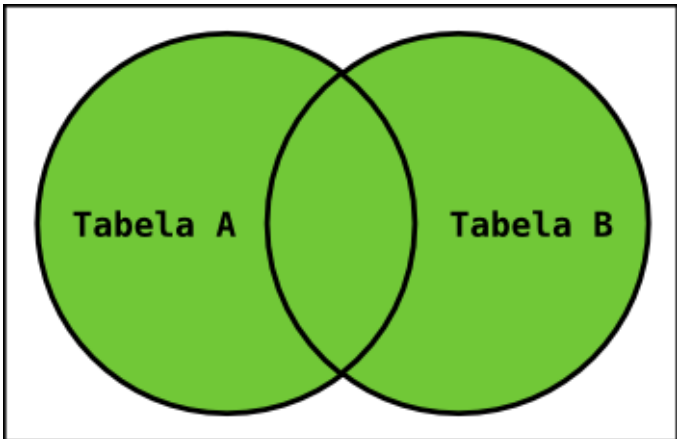
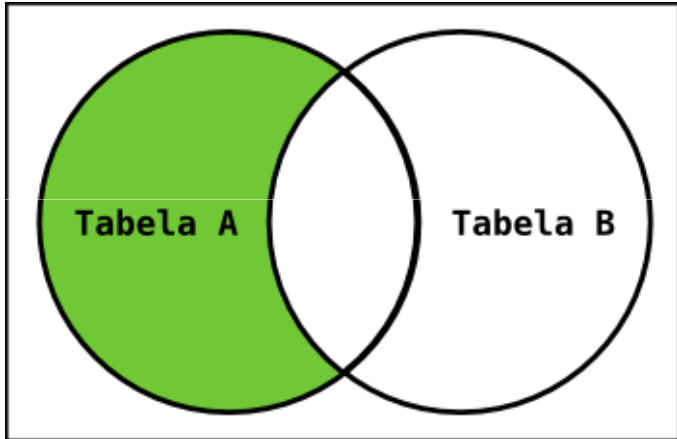
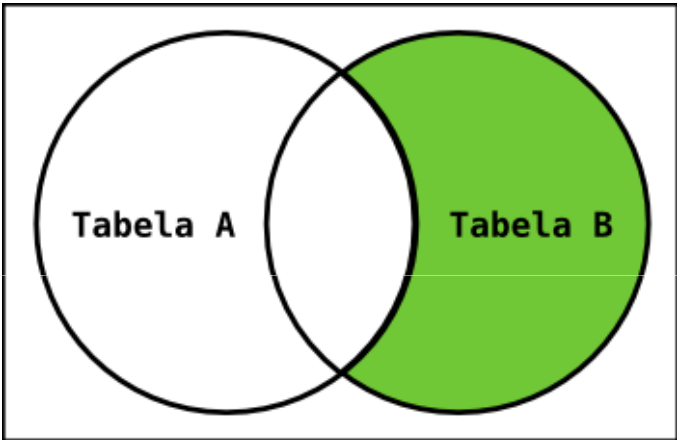
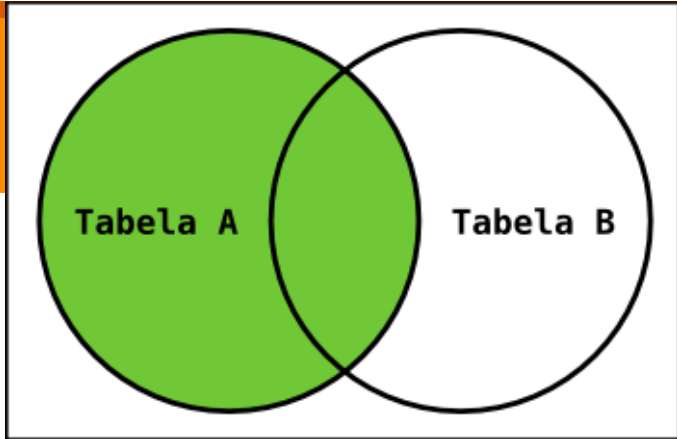
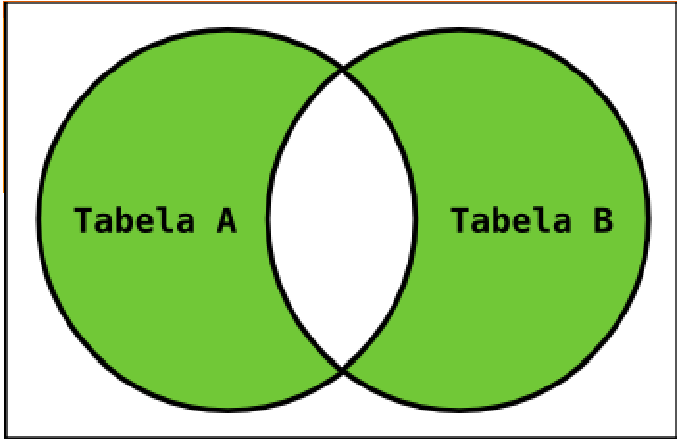


Matemática!!!

SQL foi originalmente baseada na álgebra relacional, uma forma de cálculo sobre conjuntos e relações



```
SELECT *  
FROM TabelaA  
INNER JOIN TabelaB  
ON A.id = B.id
```



Mas tem que programar?

Sim, até em seus sonhos!!!



Software para acessar dados



Programação para análise de volumes gigantes de dados

Integração de Dados



Motivação

- ❏ Web: **enorme** banco de dados
 - Sem (ou com pouca) estrutura
- ❏ **Documentos** são gerados para serem disponibilizados para leitura
- ❏ Alguns documentos foram gerados a partir de **consultas** a BD
- ❏ Dados podem ser **extraídos** das páginas web para serem utilizados por **outros programas**



Web X Banco de Dados

Onde a Web é diferente de BD:

- ❑ não tem estrutura uniforme
- ❑ não tem restrição de integridade
- ❑ não tem uma linguagem de consulta ou um modelo de dados padrão

- ❑ Mas a Web é rica em **DADOS!!!!!!**



Integração de Dados

- ❏ **Problema:** A Web contém um número **crescente** de fontes de informação **autônomas** que podem ser vistas como um grande **repositório** de dados
- ❏ **Tarefa:** Responder consultas que podem requerer **extração** e **combinação** de dados de **várias** fontes de dados na WEB



Integração de Dados

🖥️ **Exemplo:** Considere um repositório de Cinema na Internet contendo dados sobre elenco, gênero e diretores. Informações sobre críticas de filmes podem ser encontradas em outras fontes web (ex: revistas) e muitas outras fontes provêem horários dos cinemas.

Combinando dados de todas estas fontes, podemos responder a:

~ **“Quais os filmes com Julia Roberts, suas críticas e horários de exibição, em cartaz hoje a noite em Paris”**



Sistemas de Integração de Dados

Duas abordagens:

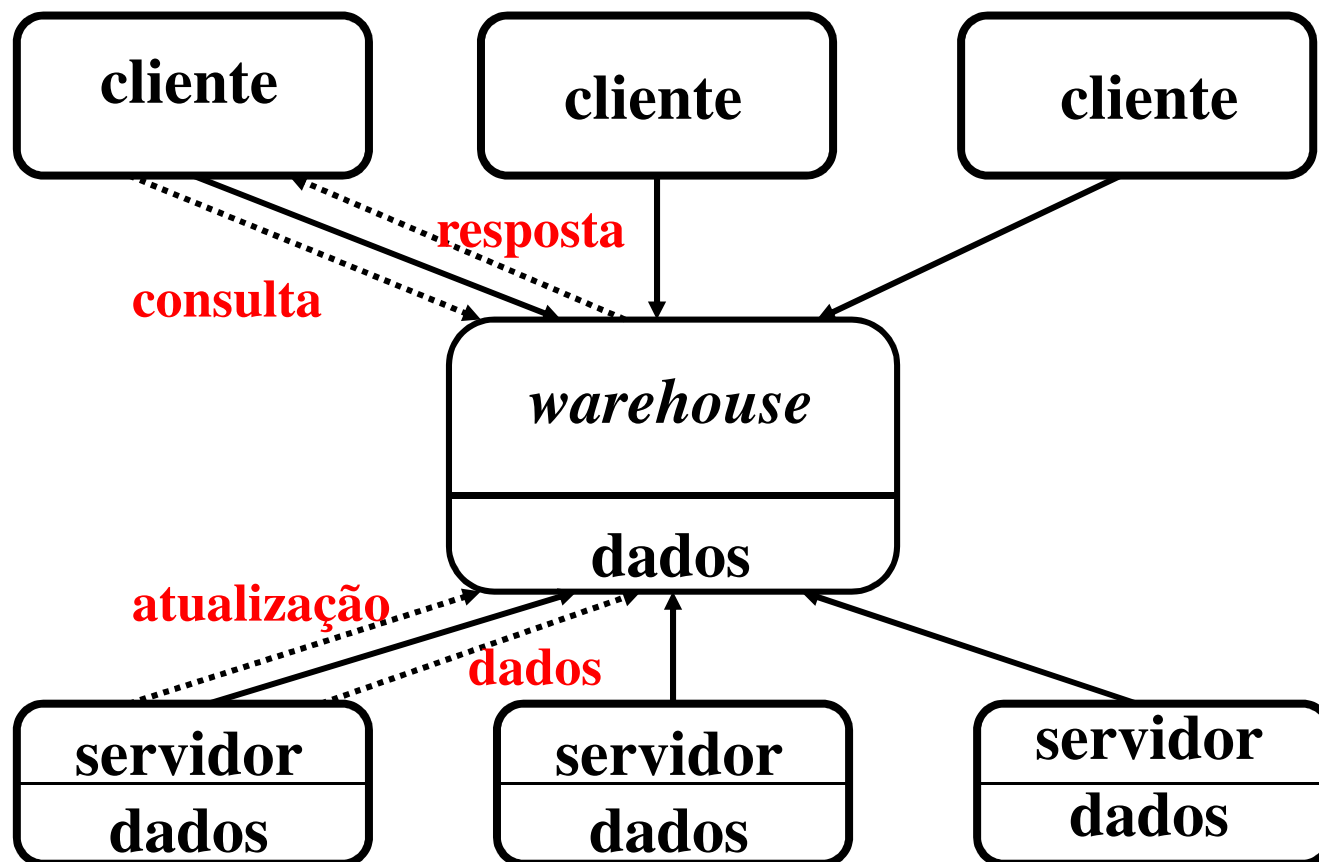
1. *Warehousing*: os dados das **diversas** fontes Web são carregados (materializados) em um repositório (***data warehouse***) e as consultas são submetidas a este repositório

Vantagem: **performance** garantida no momento da consulta

Desvantagem: atualização do repositório sempre que houver **mudança** nos dados



Arquitetura com Data Warehouse





Sistemas de Integração de Dados

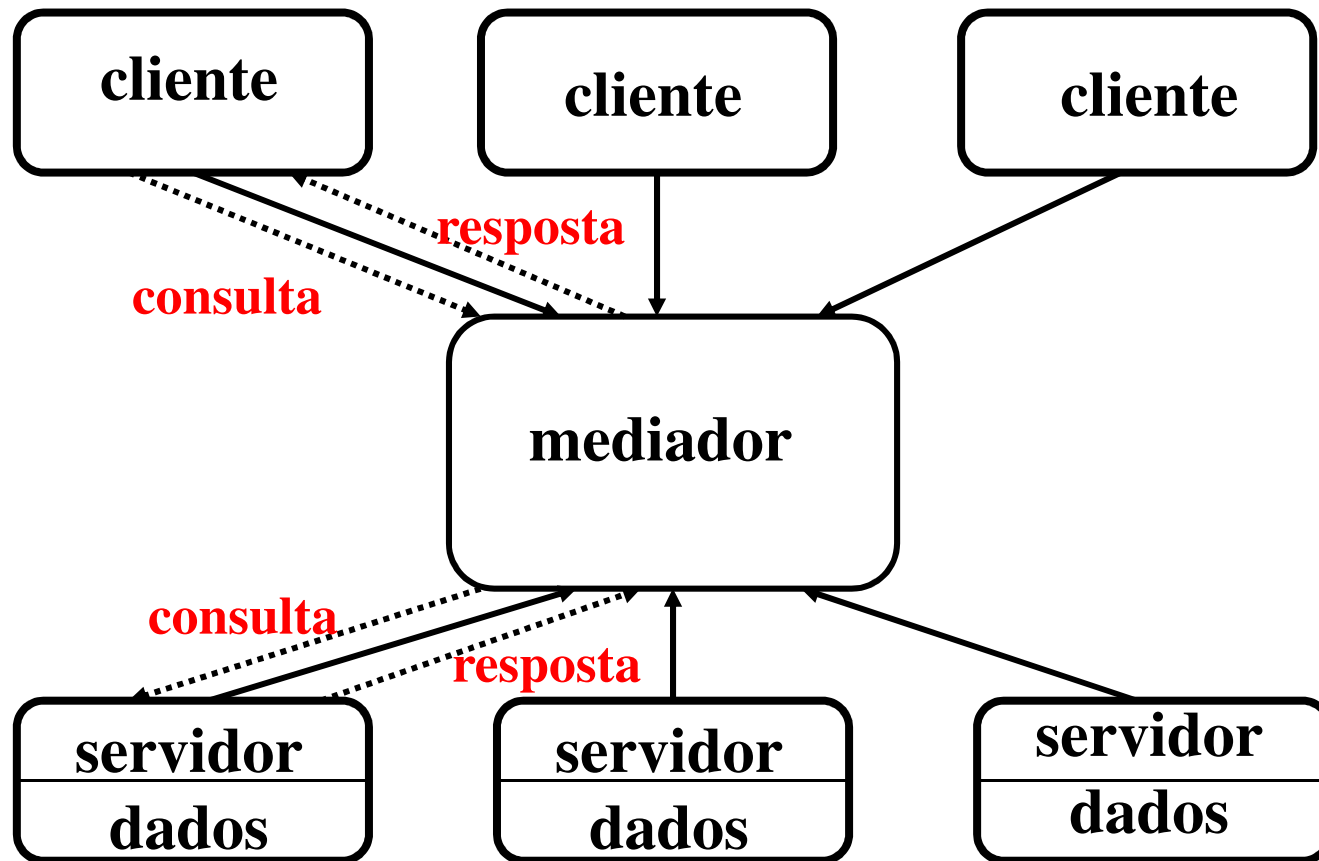
2. Enfoque Virtual: os dados são **mantidos** nas fontes Web e as consultas são submetidas diretamente às **diversas** fontes

Vantagem: os dados não são **replicados** e tem-se a garantia de estarem **atualizados** no momento da consulta

Desvantagem: como as fontes de dados são **autônomas**, são necessários métodos para otimização de consultas para garantir uma **performance** adequada



Arquitetura com Mediador





Qualidade de Dados





Motivação

- ❏ Alguns aspectos em BD como **integridade** e **segurança**, vêm sendo estudados desde a introdução da tecnologia relacional
- ❏ O conceito de *qualidade de dados* emergiu nos últimos anos



Motivação

- ❏ Qualidade da **informação** (QI) tornou-se um ponto crítico em empresas e em **pesquisas**
 - ↪ **Ainda mais com o volume/velocidade/variedade**
- ❏ A **falta** de qualidade nas informações pode impactar **severamente** a efetividade de uma organização
 - ↪ O crescimento do uso de tecnologias de **data warehouse** e acesso direto a **fontes de dados** distribuídas



Qualidade da Informação (QI)

☐ Informações de boa qualidade são informações que podem ser “usadas pelos consumidores”

↪ ***Fitness for use***

☐ Conceito **multidimensional**:

↪ Assinalar valores de escores para critérios de qualidade associados à informação



Qualidade da Informação (QI)

☐ Critérios de qualidade ou **dimensões**.

Exemplos:

↪ **Precisão**

☐ Grau de corretude e precisão que reflete o domínio da aplicação

↪ **Compleitude**

☐ Todos dados relevantes para a aplicação estão disponíveis

↪ **Atualidade**

☐ A "idade" do dado

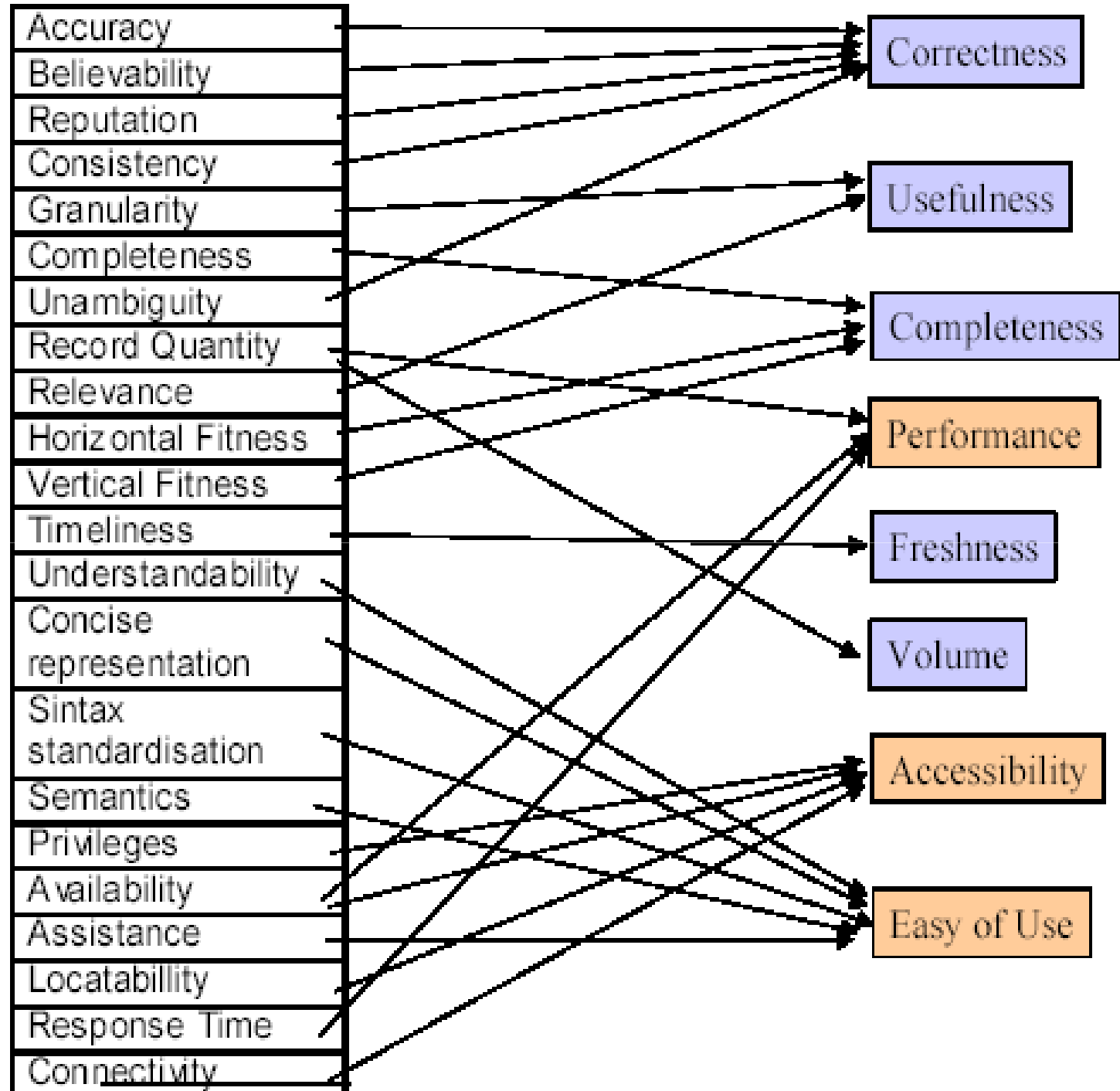
↪ **Disponibilidade**

☐ O tempo que uma fonte fica disponível

↪



Critérios de QI



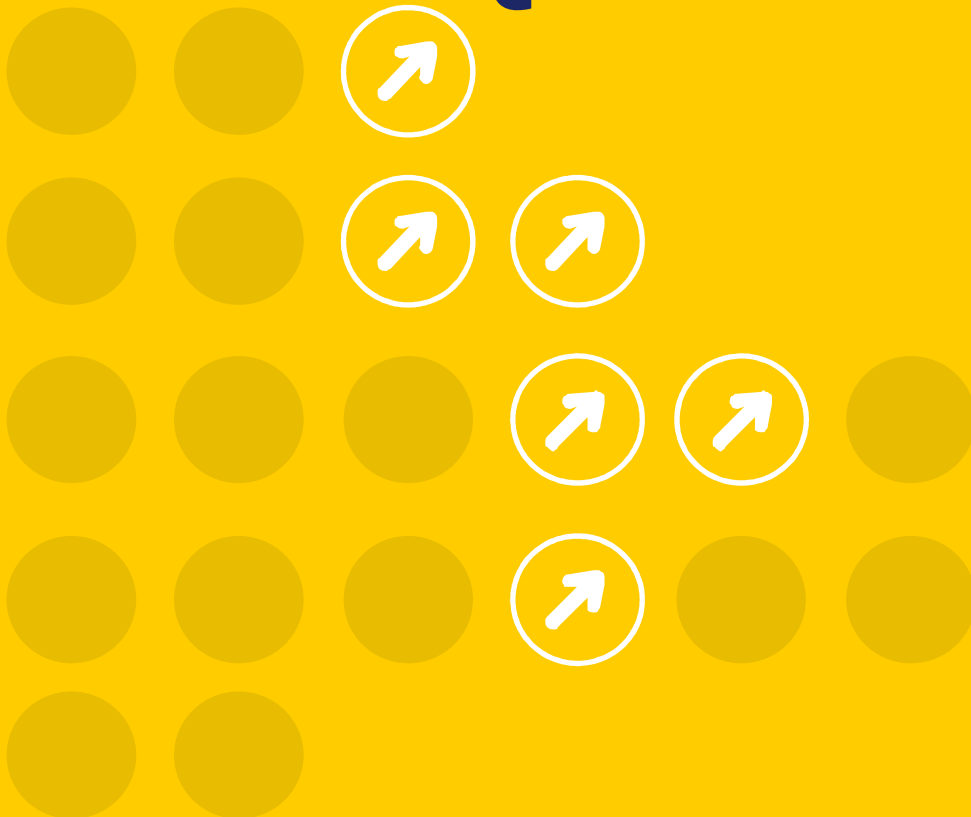


Qualidade da Informação

- ☐ Em sistemas de **Integração de Dados**
 - ↪ **Seleção** de fontes
 - ↪ **Esquemas** das fontes
 - ↪ **Esquemas** do mediador
 - ↪ Processamento de **consultas**
 - ↪ **Resultados** das consultas
 - ↪ ...



Banco de Dados, Integração e Qualidade de Dados



Ceça Moraes

cecafac@gmail.com