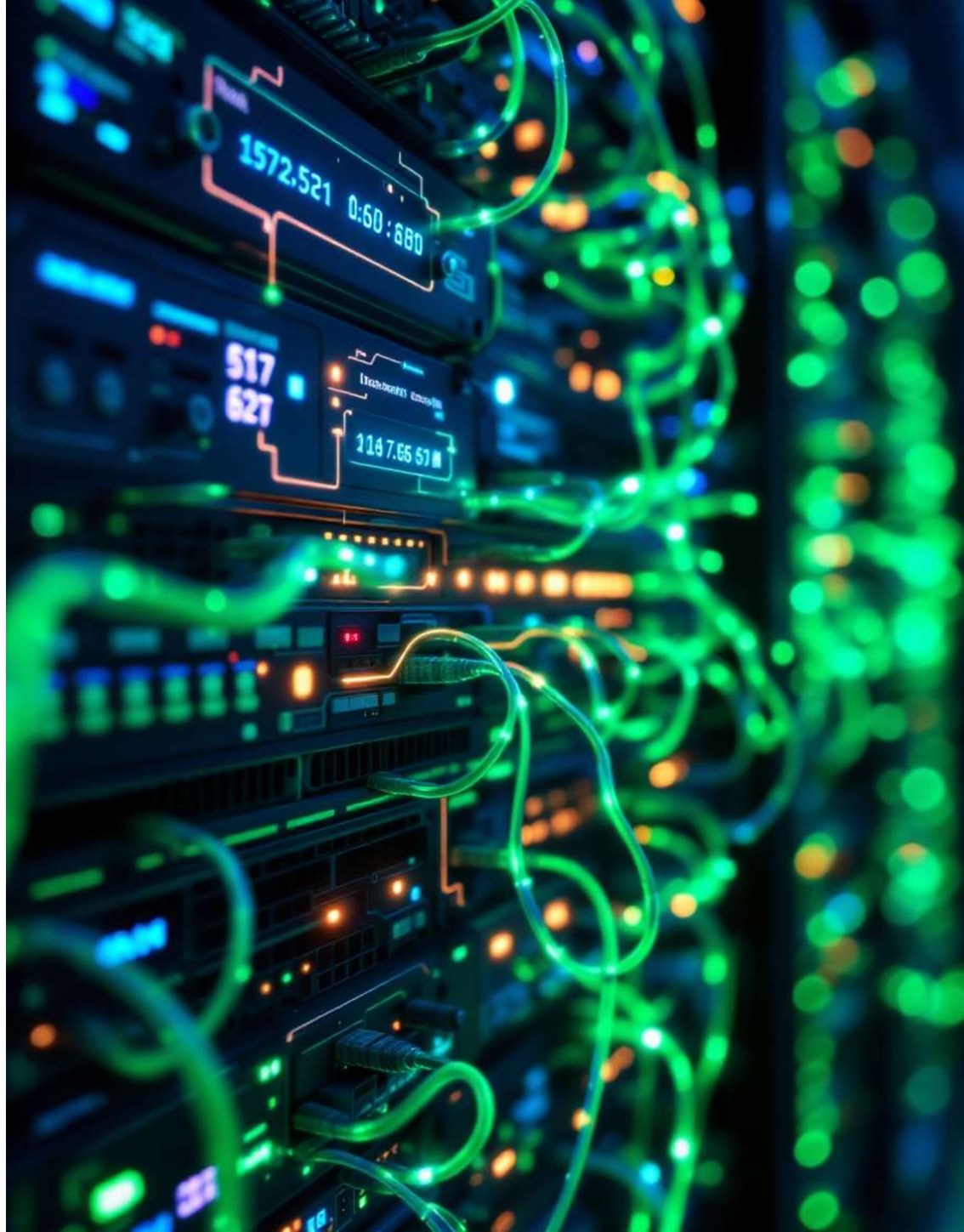
Data Mining: Descoberta de Conhecimento em Base de Dados

Data Mining é uma técnica que permite a um usuário utilizer técnicas para descobrir conhecimentos que não estão visíveis a olho nu e que poderá dar apoio na tomada de decisões importantes para uma determinada organização

Prof. Msc. Edgard Devanir Amoroso



Gestão do Conhecimento

"Disciplina que promove uma abordagem integrada para **identificar, gerenciar e compartilhar todos os recursos de informação de uma empresa**, incluindo bancos de dados, documentos, políticas e procedimentos assim como especialidades não articuladas e experiências **residentes na mente** de cada indivíduo dentro da organização."

(Gartner Group)

"Conjunto de procedimentos, infra-estrutura tecnológica, práticas e ferramentas para possibilitar **a efetiva aquisição, organização e distribuição de informações relevantes**, para as pessoas certas no tempo certo, de modo a capacitá-las a contribuir na realização dos objetivos do negócio através de ações eficazes."

(Ernest & Young)

Gestão do Conhecimento

Aprendizagem Organizacional

Habilidade de uma empresa **aprender** e rapidamente **traduzir em ações** o conhecimento como um meio eficaz de atingir vantagem competitiva.

- **O que uma empresa conhece, como usa o que conhece e com que rapidez pode aprender algo novo.**

Gestão do Conhecimento

Objetivo

Melhorar a qualidade da **tomada de decisão** em todos os níveis da organização, através do aumento do **acesso às informações** e da **redução do problema** de sobrecarga de informações.

Gestão do Conhecimento

Caminho do conhecimento

Gestão do Conhecimento

O que são dados?

- As **evidências mais básicas de uma investigação**, aqueles aspectos do fenômeno sendo estudado que um determinado investigador pôde captar e registrar
- Correspondem as **observações consideradas diretas**, ou seja, com relativamente pouca elaboração ou tratamento
- Uma vez coletados, são compreendidos como um reflexo razoavelmente confiável dos acontecimentos (fatos) concretos
- **Exemplos**
  - Registros de ligações telefônicas com seu tempo de ligação, origem e destino
  - Corretores incluem características de clientes em seus seguros
  - Atendentes incluem produtos vendidos em notas fiscais
  - Médicos incluem medidas anatômicas, bioquímicas e laboratoriais em geral, bem como os resultados de exames clínicos e de testes de função.

Gestão do Conhecimento

O que é informação?

- É o resultado ...
  - ... de uma **organização, transformação e/ou análise de dados**
  - ... do tratamento de um conjunto de dados de modo a **produzir um significado**
- Válida apenas quando o tratamento é feito de forma **científica**, **não-induzida** e obedecendo a **regras** da boa teoria quantitativa e qualitativa
- Exemplos
  - Informações telefônicas de tráfego, receitas, custos e demandas não atendidas
  - Informações de sinistros em seguros classificados por tipo de evento
  - Informações médicas que podem
    - Incluir ocorrências, incidências, riscos e associações entre eventos, assim como resumos e descrições de grupos e populações.
    - Ser produzidas por cruzamentos de dados médicos e não-médicos

Gestão do Conhecimento

O que é conhecimento?

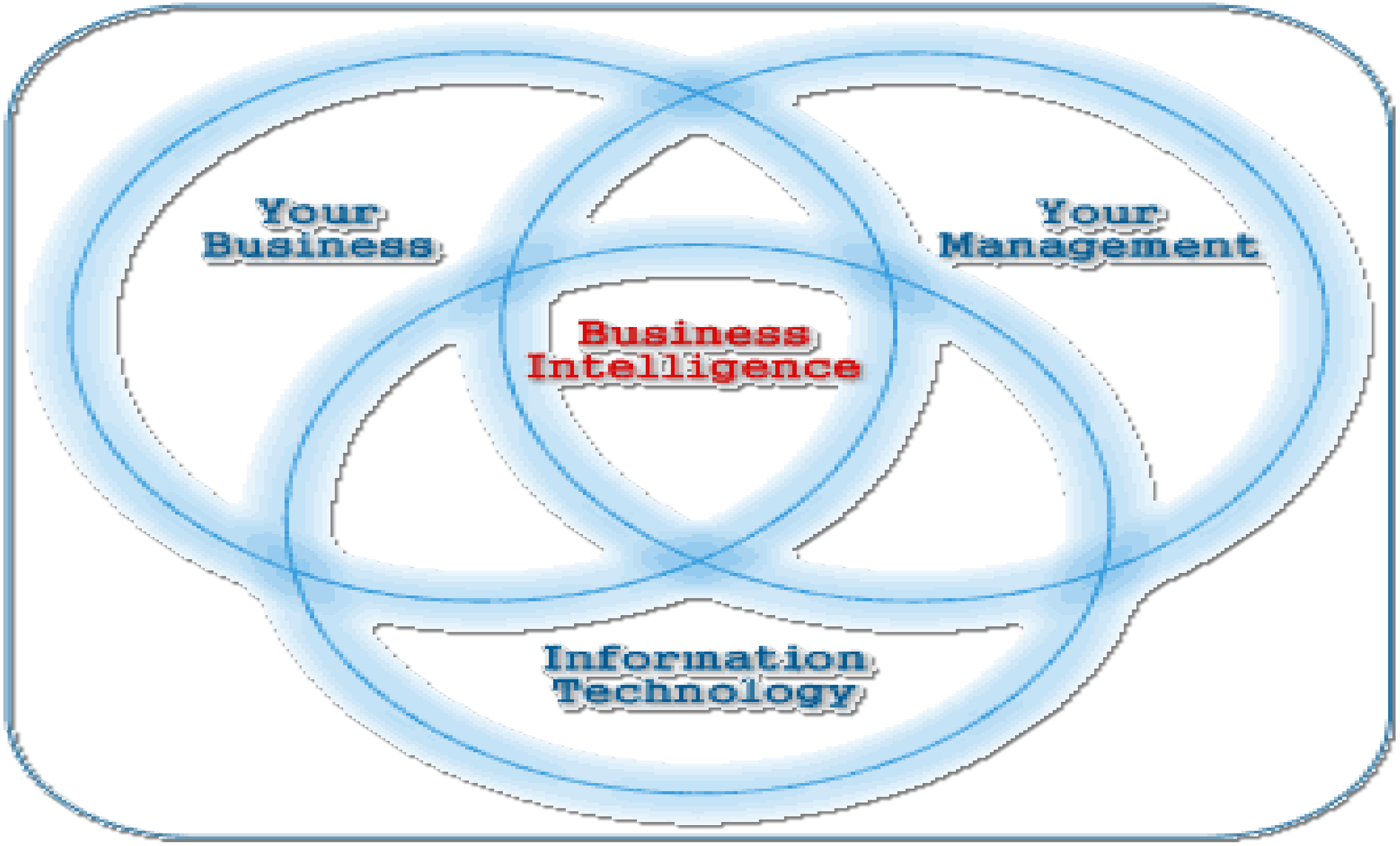
- É um conjunto de **argumentos e explicações** que interpretam um conjunto de informações
- Trata-se de **conceitos** e **argumentos lógicos** essencialmente abstratos que interligam e dão significado a fatos concretos
- Envolve **hipóteses**, **teorias**, **modelos** e **leis**
- **Exemplos**
  - Identificação de clientes da empresas de telefonia móvel que tendem a deixar empresa
  - Identificar possíveis fraudes em prêmios de seguros
  - Identificação da etiologias (causa da doença) e de mecanismos patogênicos, além da elaboração de modelos do funcionamento de sistemas fisiológicos

Gestão do Conhecimento

Relação entre os níveis do saber

- O processo de construção de conhecimento científico envolve os dados, os quais representam a "matéria prima“ bruta, a partir dos quais as operações lógicas criam informações e, finalmente, estas últimas são interpretadas para gerar conhecimento
- É a ponte entre o empírico e o teórico, com o fenômeno gerando dados, os dados gerando informações e as informações gerando ou confirmando um conhecimento abstrato
- **Exemplo**
  - Pode-se automatizar o acompanhamento do cliente de uma empresa de telefonia móvel, seguradora, banco, plano de saúde etc, desde o momento de sua entrada na empresa, monitorando-o individualmente ou em grupos específicos e até compará-lo com outros clientes ou grupos similares

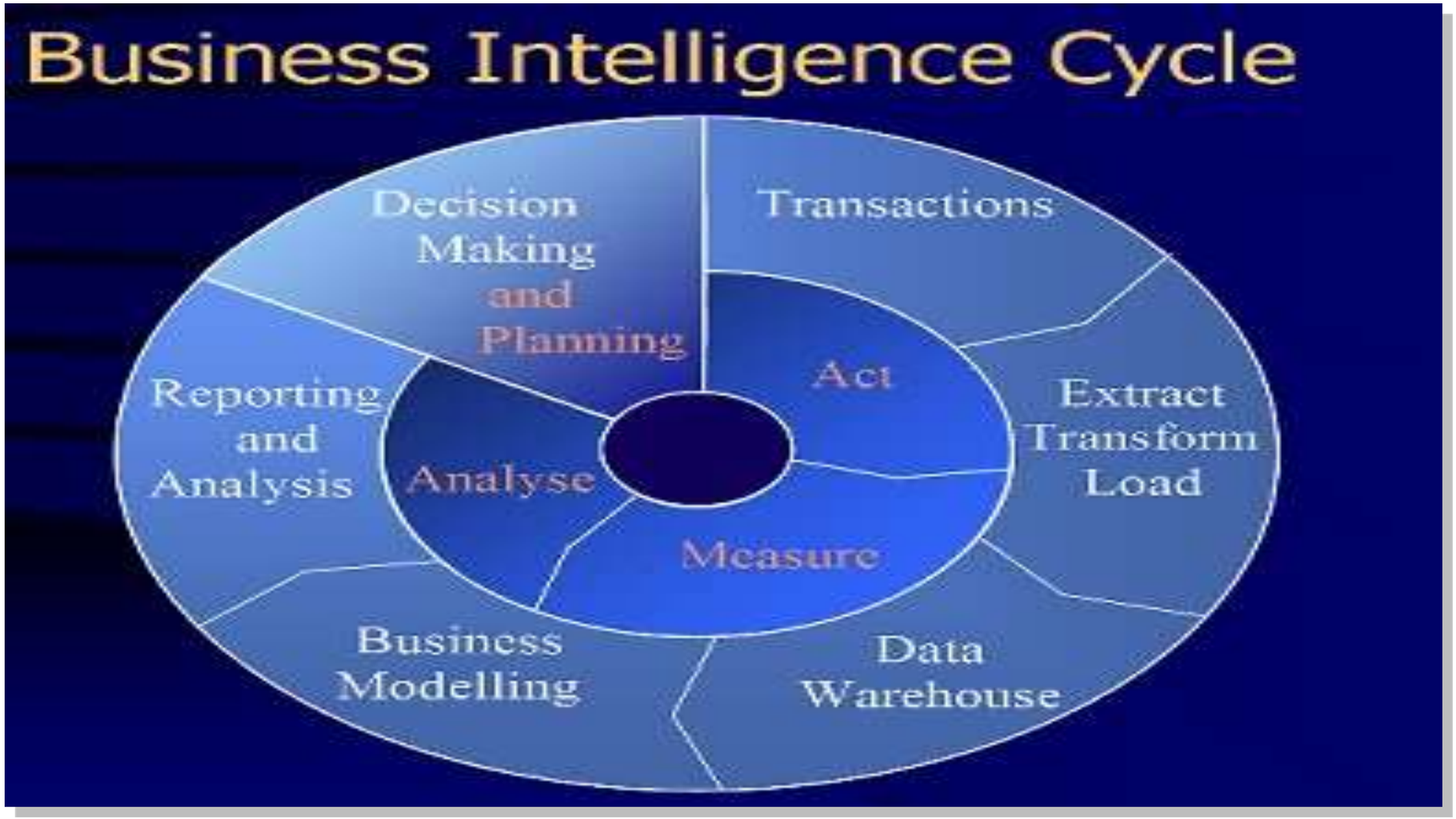
Business Intelligence



Business Intelligence

- **Componentes** para um sistema de BI
  - Database Marketing
  - CRM (Customer Relationship Management)
  - Data Warehousing
    - Data Warehouse
    - Data Mart
    - Olap (On-Line Analytical Processing)
  - Data Mining
  - Balanced Scorecard

Business Intelligence



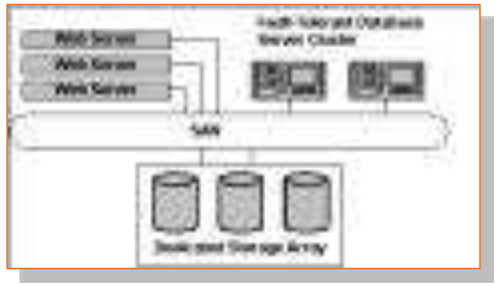
Onde está o Conhecimento Organizacional?

Onde está o Conhecimento Organizacional?

O Conhecimento Organizacional está em ...

- **Pessoas**
  - Experiências
  - Casos
  - Rotinas
  - Observações
  - Requisitos
  - Códigos
  - Especificações
  - Mensagens

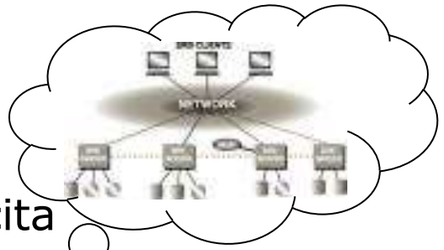
- Bases de Dados
- Documentos
- Correspondências
- Arquivos
- Livros
- Filmes
- Textos
- Planilhas
- ....



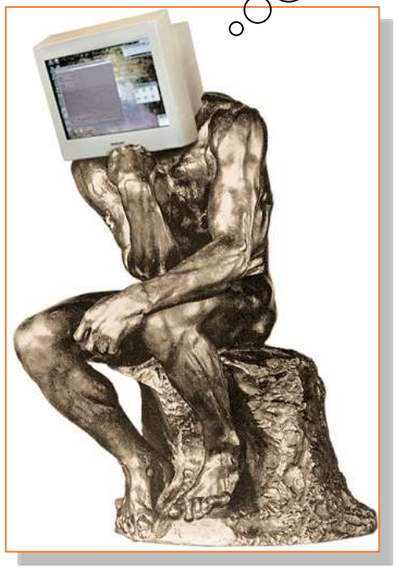
Conhecimento & Gestão do Conhecimento

Conhecimento segundo a Gestão do Conhecimento

- Tácito, “na cabeça das pessoas”
- Não permite representação
- Difícil de explicar e se elicitar
- Se torna dados e informação quando assume forma explícita



➔ **Explicação**
Elicitação ➔

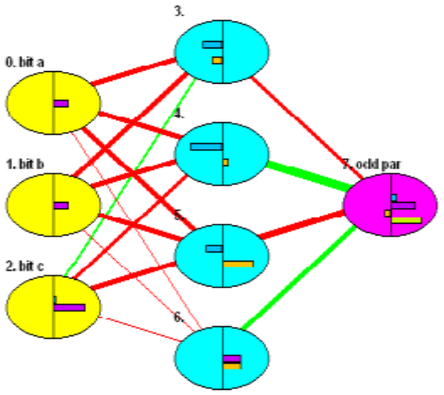


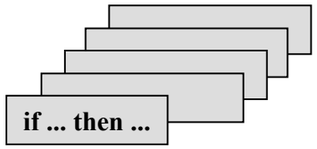
Conhecimento & Inteligência Artificial

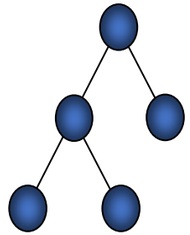
- Representação do Conhecimento
- Sucessos em diversos formas de representação

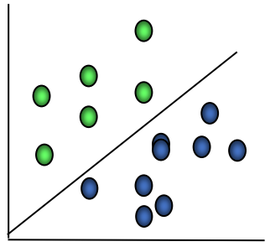
Ex. de formas de representação:

- Simbólicas
  - Frames e Redes Semânticas
  - Regras de Produção
  - Árvores de Decisão
- Conexionistas (RNA´s)
  - Classificação (Supervisionados)
  - Agrupamento - Clustering (Não Supervisionados)









Processo de Transferência do Conhecimento

Espiral de criação do Conhecimento *(Nonaka & Takeuchi)*

Externalização do Conhecimento

"*Externalização* é um processo de articulação do conhecimento tácito em conceitos explícitos. O tácito se torna explícito expresso na forma de metáforas, analogias, conceitos, hipóteses ou modelos."

(Nonaka & Takeuchi)

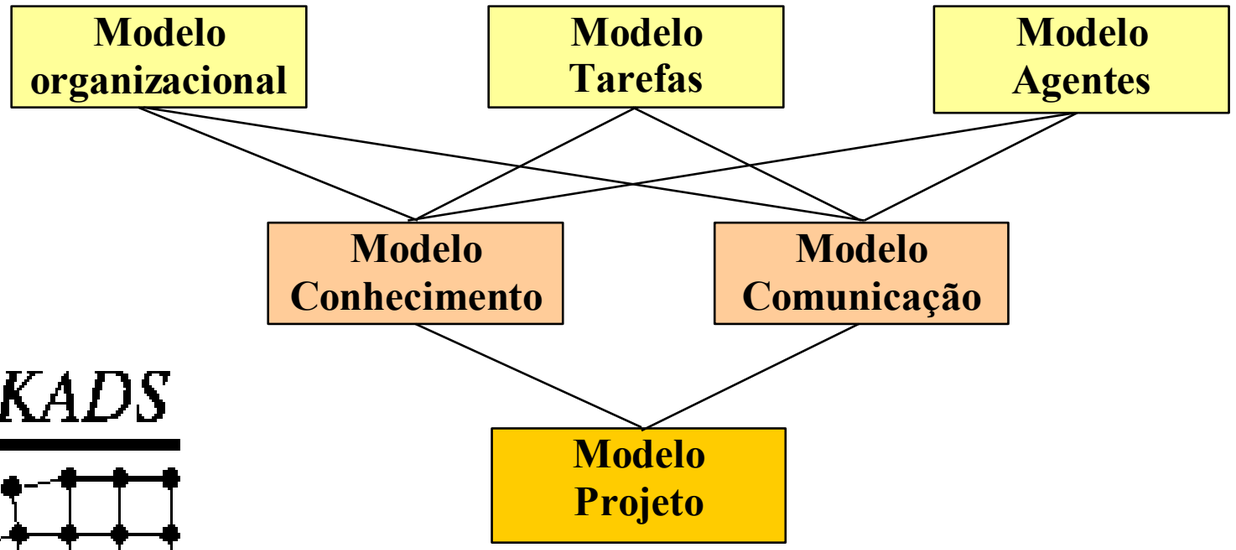
Metodologias

- Engenharia do Conhecimento - Metodologia Commonkads
  - Empírico
- Reconhecimento de Padrões - Metodologia CRISP-DM
  - Analítico

Externalização do Conhecimento

Engenharia do Conhecimento

- **Commonkads** é um recurso útil para externalização do conhecimento, ou seja, conversão do conhecimento tácito em explícito.

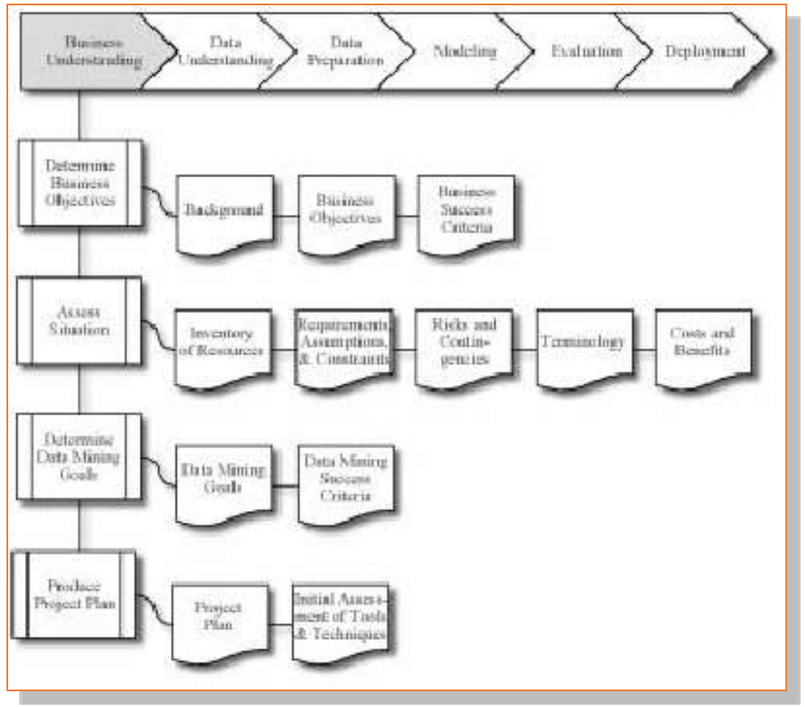


CommonKADS

Externalização do Conhecimento

Reconhecimento de Padrões

- **CRISP-DM** é uma metodologia que busca orientar o processo de mineração de dados em bases de dados



Externalização do Conhecimento

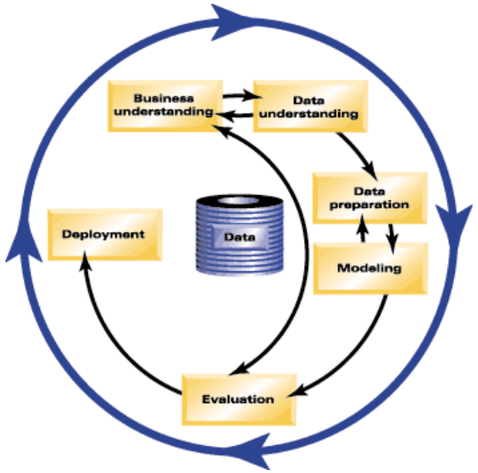
***Knowledge Discovery in databases* (KDD ou DCBD)**

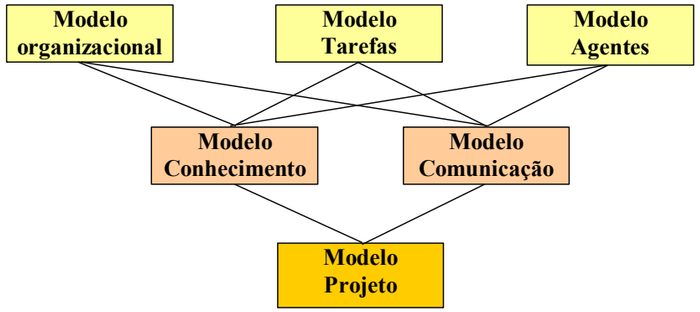
Termo utilizado por pesquisadores de Inteligência Artificial para o processo para encontrar conhecimento em dados (Aprendizagem de Máquina)

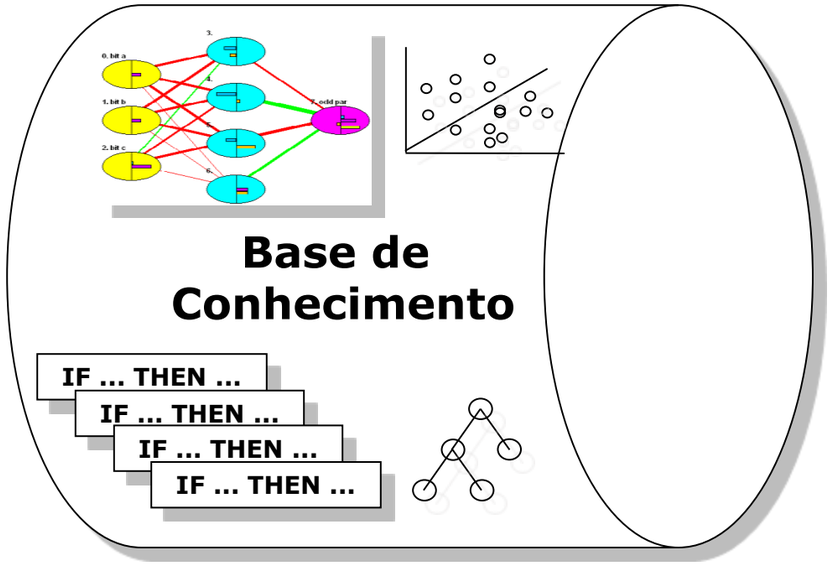
Mineração de dados (*Data Mining*)

Termo utilizado por estatísticos, analistas de dados e pela comunidade de gestão de sistemas de informação.

Externalização do Conhecimento

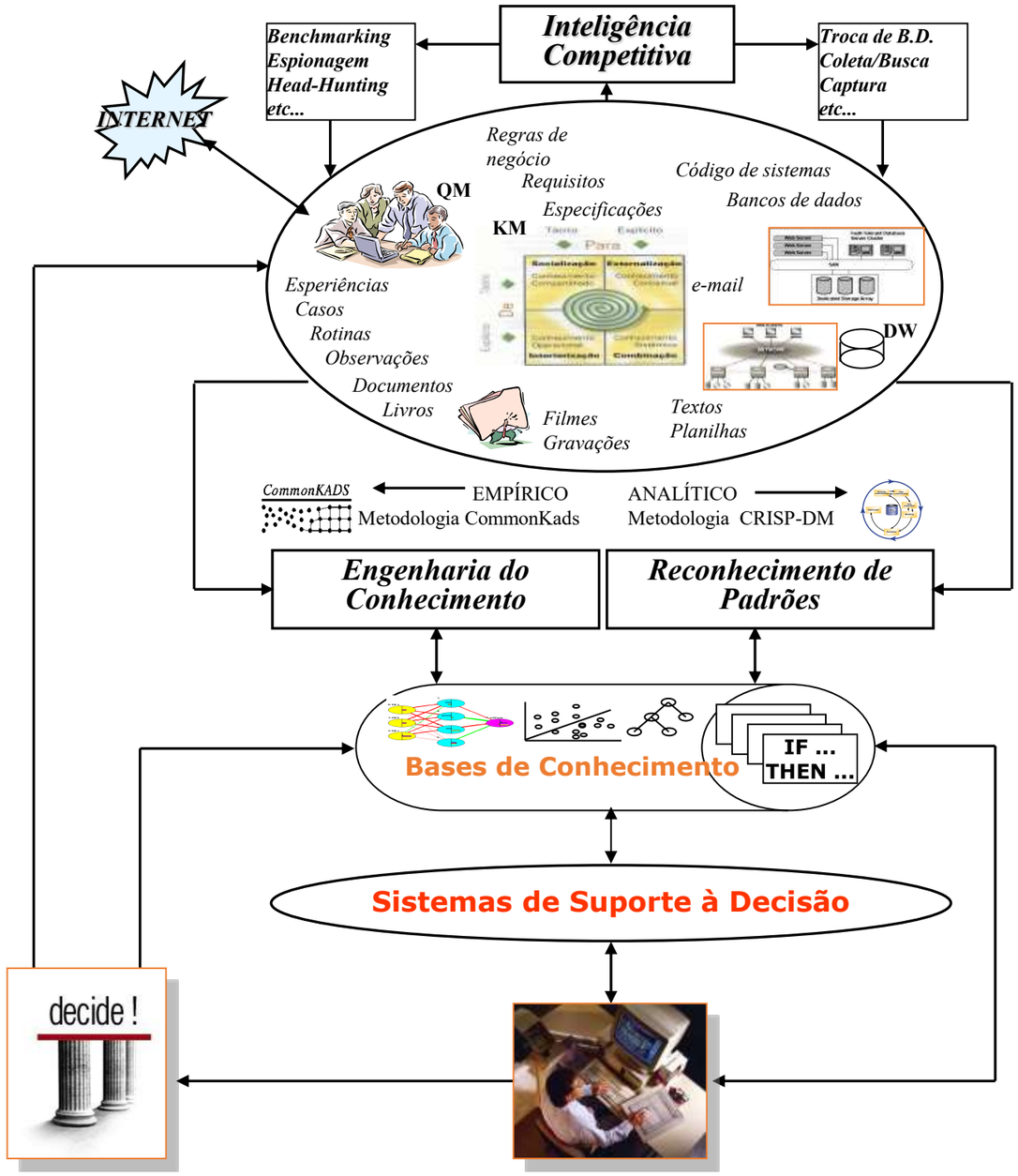






Conhecimento Organizacional

Modelo para Descoberta e Aquisição do Conhecimento Organizacional

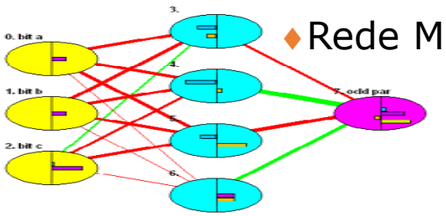


Bases de Conhecimento

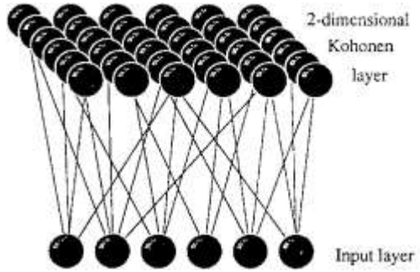
♦ Exemplos de Bases de conhecimento

• Redes Neurais Artificiais

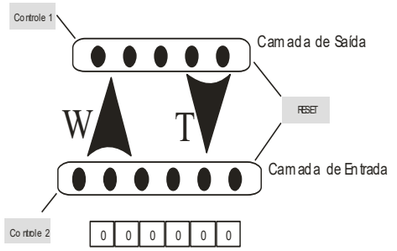
♦Rede MLP



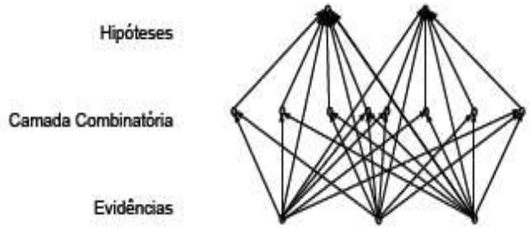
♦Rede SOM-Kohonen



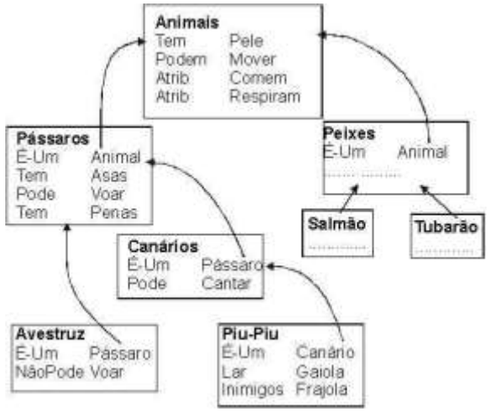
♦Rede ART1



♦CNM – Modelo Híbrido Neuro Simbólico



♦Frames e Redes Semânticas



Regras de Produção

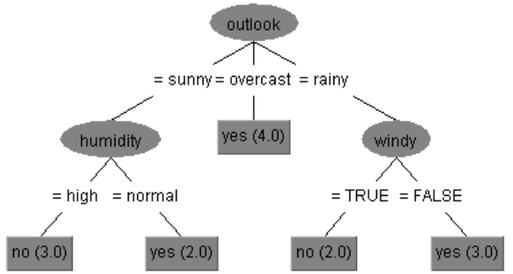
IF ... THEN ...

IF ... THEN ...

IF ... THEN ...

IF ... THEN ...

Arvores de Decisão



KDD – Knowledge Discovery Database – Descoberta de Conhecimento

– KDD é um processo que envolve um conjunto de etapas para automação da identificação e do reconhecimento de padrões em bases de dados.

– KDD como também é conhecido, é ao processo de extração não trivial, previamente desconhecida, de informações potencialmente úteis dos dados.

– Isto é, através da utilização de técnicas computacionais é possível extrair de um repositório de dados, informações que não estão explicitamente representadas nele, impossíveis de serem obtidas através de técnicas estruturadas como Queries e OLAP.

Datamining (Mineração de Dados) - Definições

- **Mining**
  - **Datamining**
    - é uma etapa do processo de descoberta de conhecimento onde se aplicam os métodos e algoritmos específicos para busca dos padrões implícitos nestes dados.
  - **WebMining**
    - Descoberta de conhecimento na WEB
  - **TextMining**
    - Descoberta de conhecimento em textos e arquivos não estruturados.

Datamining (Mineração de Dados) - Definições

O termo Datamining foi estendido além seus limites para aplicar a qualquer forma de análise de dados. Algumas das numerosas definições de Datamining, ou KDD são:

– "Datamining, ou KDD como também é conhecido, é a extração não trivial, previamente desconhecida, de informações potencialmente úteis dos dados. Isto engloba várias técnicas de aproximações diferentes, como clustering, sumarização de dados, regras de aprendizado de classificação, análise de mudanças e detecção de anomalias".

  William J Frawley, Gregory Piatetsky-Shapiro e Christopher J Matheus

Datamining (Mineração de Dados) - Definições

- "Datamining é a procura pôr relações e padrões globais que existem *em **grandes*** bancos de dados mas estão escondidos na vasta quantia de dados, como uma relação entre os dados de um paciente e seu diagnostico médico. Estas relações representam valioso conhecimento sobre o banco de dados e os objetos pertencentes a este e se o banco de dados é um espelho fiel do mundo real registrado pelo banco de dados."
  Marcel Holshemier & Arno Siebes (1994)

  Qual o significado do Grandes????

Onde aplicar KDD

- Regras de atribuição de crédito e análise de Risco
- Perfil de Consumidor
- Associações e Perfil de Clientes
- Fraudes em finanças (Bancos e Cartões de Crédito)
- Tempo de Internamento de Doentes em Hospital
- Identificar terapias e medicamentos
- Perfil de pessoas
- Séries Temporais
- Marketing
- Etc...

Técnicas de Data Mining

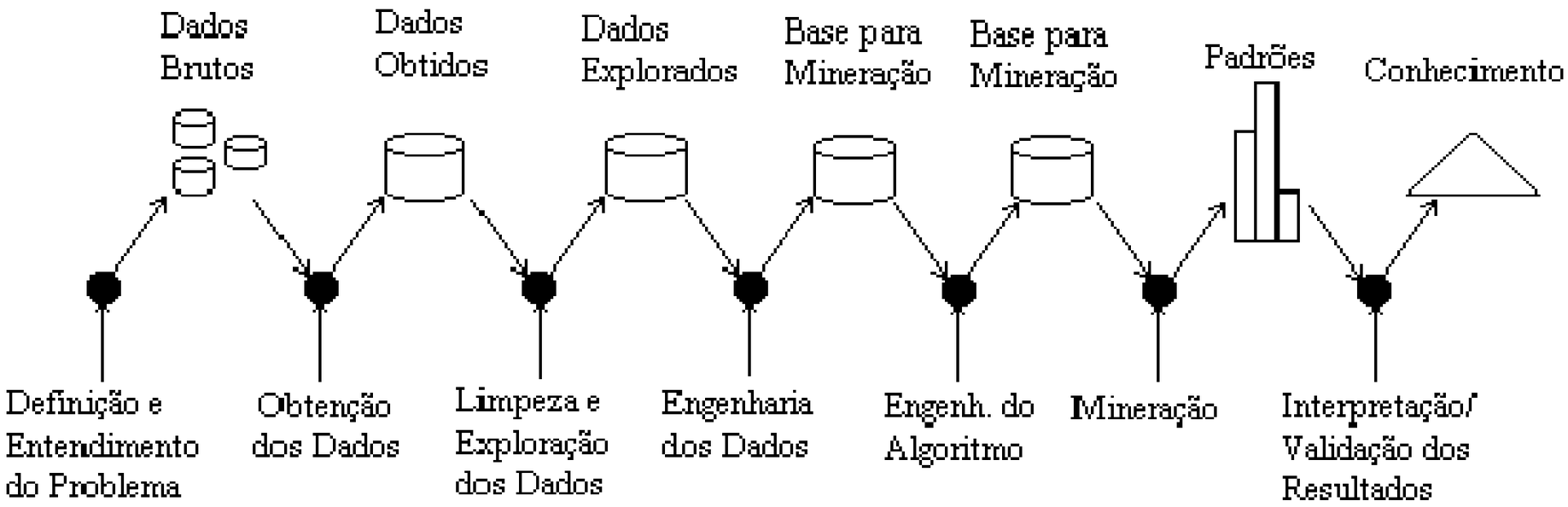
- Principais Técnicas

  – Classificação / Regressão (RNA´s)
  – Agrupamentos / Clustering (RNA´s e Métodos Estatísticos)
  – Regras de Associação (Métodos Indutivos)
  – Predição / Séries (RNA´s)

Abordagem Metodológica de KDD (DCBD)

As fases do processo

O processo de KDD pode ser apresentado como uma seqüência que se dá de forma iterativa, no sentido em que pode-se avançar nas etapas e posteriormente voltar a uma que já tenha sido executada, e interativa pelo fato de se basear intensamente nas entradas e respostas providas pelo usuário.



O processo completo de KDD, consta das etapas de Definição e Entendimento do Problema, Obtenção dos Dados, Limpeza e Exploração dos Dados, Engenharia dos Dados, Engenharia do Algoritmo, Mineração, Interpretação e Validação dos Resultados.

Etapas de KDD

- KDD é um processo amplo que consiste das seguintes etapas:

  - **Seleção**: ocorre a seleção dos atributos que interessam ao usuário.
  - **Integração dos dados**: ocorre a combinação de diferentes fontes de dados produzindo um único repositório de dados (Data Warehouse).
  - **Limpeza de dados**: ocorre a eliminação de ruídos e dados inconsistentes.
  - **Transformação dos dados**: os dados são transformados num formato apropriado para aplicação de algoritmos de mineração.
  - **Mineração**: consiste na aplicação de técnicas inteligentes a fim de extrair os padrões de interesse.
  - **Avaliação ou pós-processamento**: ocorre a identificação dos padrões interessantes de acordo com algum critério do usuário.
  - **Visualização dos resultados**: ocorre a utilização de técnicas de representação do conhecimento com o objetivo de apresentar ao usuário o conhecimento minerado.

Referências Bibliográficas

- Ye, Ye. The handbook of data mining. Lawrence Erlbaum Associates, Publishers, 2003. New Jersey. London.

Text Mining: Transformando Dados Textuais em Conhecimento

Bem-vindos à era do conhecimento a partir de dados! Nesta apresentação, vamos desvendar o mundo do Text Mining, uma disciplina essencial para converter a vasta quantidade de texto disponível no ambiente digital – desde redes sociais a documentos corporativos – em insights acionáveis. Nosso objetivo é mostrar como o processamento de textos pode gerar valor significativo em diversos setores.

Prof. Msc. Edgard Devanir Amoroso

O Que É Text Mining?

Text Mining é o processo de extrair informações valiosas e padrões relevantes de grandes volumes de dados textuais que, na sua forma original, são não estruturados ou semi-estruturados. É a ponte que conecta o texto bruto ao conhecimento estruturado.

Definição Ampliada

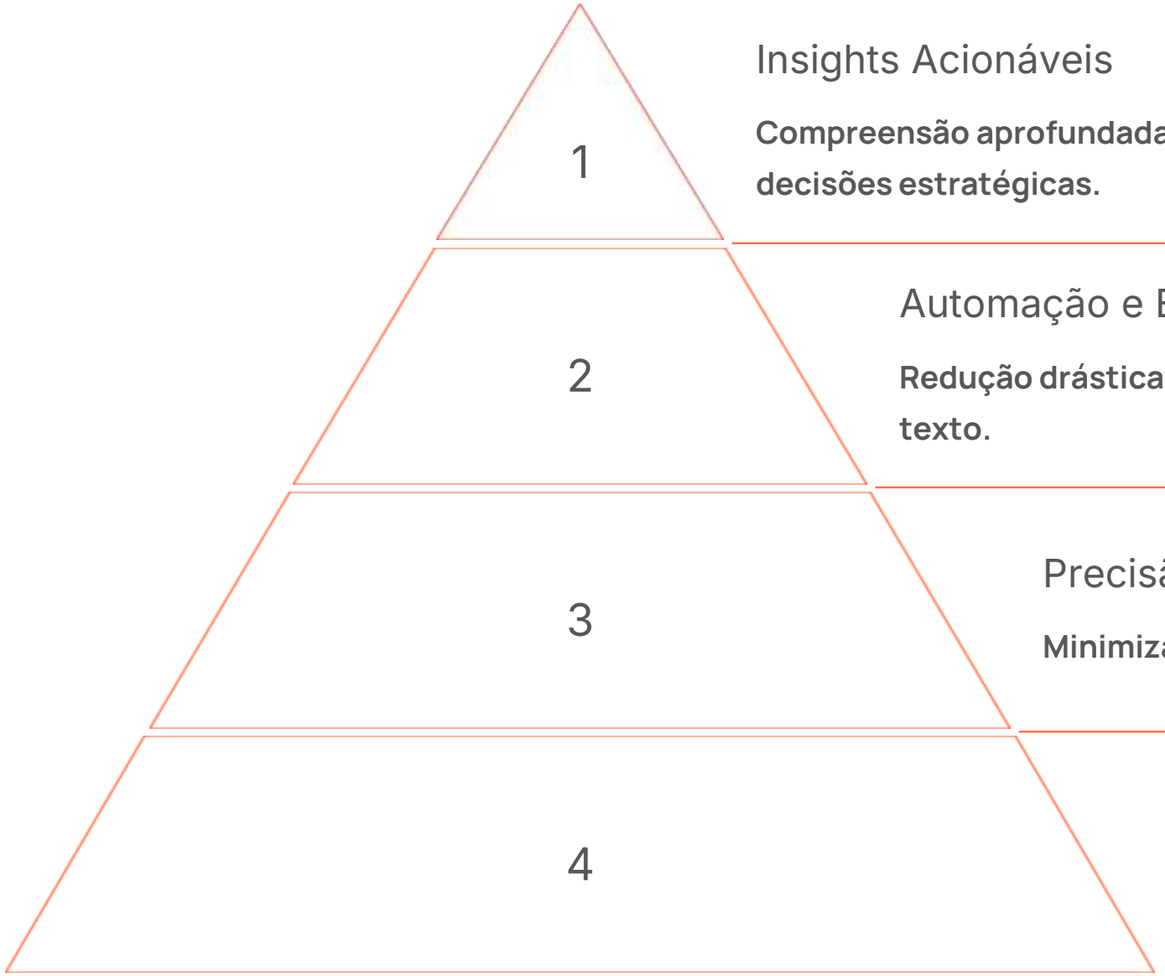
Utiliza uma combinação de técnicas de Processamento de Linguagem Natural (PLN), estatística e aprendizado de máquina para transformar o texto em um formato que pode ser analisado quantitativamente. Isso permite a identificação de tendências, a classificação de conteúdo e a descoberta de relações ocultas nos dados.

Exemplo Prático

Imagine uma empresa de e-commerce que recebe milhares de comentários de clientes diariamente. Através do Text Mining, é possível analisar esses comentários em larga escala para identificar os principais pontos fortes do produto, áreas de melhoria e até mesmo detectar sentimentos positivos ou negativos predominantes.

Importância do Text Mining na Era dos Dados

O volume de dados textuais cresce exponencialmente, tornando a análise manual inviável. O Text Mining surge como uma solução para essa avalanche de informações, transformando-a em uma vantagem competitiva.



Insights Acionáveis

Compreensão aprofundada do mercado e do comportamento do consumidor para tomada de decisões estratégicas.

Automação e Eficiência

Redução drástica do tempo e dos recursos necessários para analisar grandes volumes de texto.

Precisão Aprimorada

Minimização de erros humanos e aumento da consistência nas análises.

Vantagem Competitiva

Identificação de oportunidades e ameaças que passariam despercebidas.

Etapas Fundamentais do Processo de Text Mining

O processo de mineração de texto é um ciclo contínuo que envolve várias etapas, desde a preparação dos dados até a apresentação dos resultados.

1

1. Coleta de Dados

Obtenção de textos de diversas fontes, como websites, bancos de dados corporativos, redes sociais e documentos variados. Essa etapa define a qualidade e a abrangência da análise.

2

2. Pré-processamento

Essencial para transformar dados brutos em um formato legível para máquinas. Inclui limpeza (remoção de ruídos, caracteres especiais, stop words), tokenização (divisão em palavras/frases), normalização (stemming, lematização) e vetorização (transformação em representações numéricas, como TF-IDF).

3

3. Análise e Extração

Aplicação de algoritmos de aprendizado de máquina e estatística, como classificação, clustering e modelagem de tópicos, para identificar padrões, temas e informações relevantes.

4

4. Visualização

Apresentação dos insights obtidos de forma clara e intuitiva, utilizando gráficos, nuvens de palavras e dashboards interativos para facilitar a compreensão e a tomada de decisão.

Técnicas Essenciais de Pré-processamento

O pré-processamento é a base para qualquer análise de Text Mining, garantindo que os dados estejam limpos e prontos para extrair informações significativas.

Limpeza e Normalização

- Remoção de Stop Words: Elimina palavras comuns (ex: "e", "de", "o") que não agregam valor analítico, reduzindo o volume de dados e o ruído.
- Stemming: Reduz palavras à sua raiz morfológica. Ex: "correndo", "corri", "corre" se tornam "corr". Útil para agrupar variações da mesma palavra.
- Lematização: Transforma palavras flexionadas em sua forma base ou lema, considerando o contexto. Ex: "melhores" se torna "bom"; "amando" se torna "amar". É mais sofisticada que o stemming.

Estruturação e Vetorização

- Tokenização: Quebra o texto em unidades menores, como palavras (tokens) ou frases. É o primeiro passo para organizar o texto.
- Vetorização: Converte os tokens em representações numéricas, pois algoritmos de Machine Learning operam com números. As técnicas mais comuns incluem:
  - TF-IDF (Term Frequency-Inverse Document Frequency): Pondera a importância de uma palavra em um documento em relação a todo o corpus.
  - Word Embeddings: Representações densas de palavras (ex: Word2Vec, GloVe) que capturam o significado semântico e as relações entre as palavras.

Principais Algoritmos e Métodos em Text Mining

Após o pré-processamento, diversos algoritmos são aplicados para extrair padrões e insights do texto, cada um com uma finalidade específica.

Classificação de Textos

Atribui categorias ou rótulos a documentos com base em seu conteúdo. Exemplo: Filtrar e-mails para identificar spam ou categorizar notícias por assunto (esportes, política, etc.). Algoritmos: Naive Bayes (simples e eficaz), Support Vector Machines (SVM) e Redes Neurais (especialmente para grandes volumes e alta complexidade).

Clustering

Agrupa textos semelhantes sem rótulos predefinidos. Útil para descobrir estruturas naturais nos dados. Exemplo: Organizar automaticamente artigos científicos por temas emergentes ou agrupar avaliações de produtos com sentimentos similares. Algoritmos: K-means (simples e rápido) e DBSCAN (identifica clusters de densidade arbitrária).

Modelos de Tópicos

Identificam os temas subjacentes que permeiam uma coleção de documentos. Exemplo: Analisar discursos políticos para descobrir os tópicos mais abordados ou examinar fóruns online para identificar as principais discussões dos usuários. Algoritmo: LDA (Latent Dirichlet Allocation) é amplamente usado para essa finalidade.

Principais Aplicações do Text Mining

A capacidade de extrair valor de dados textuais abriu um leque vasto de aplicações em diversas indústrias e setores.

Análise de Sentimento

Avalia o tom emocional de textos (positivo, negativo, neutro). Crucial para entender a percepção de marca, feedback de clientes e o humor das redes sociais.

Monitoramento de Redes Sociais

Identifica tendências, detecta crises de imagem em tempo real, mede a performance de campanhas e compreende a opinião pública sobre produtos ou serviços.

Resumo Automático

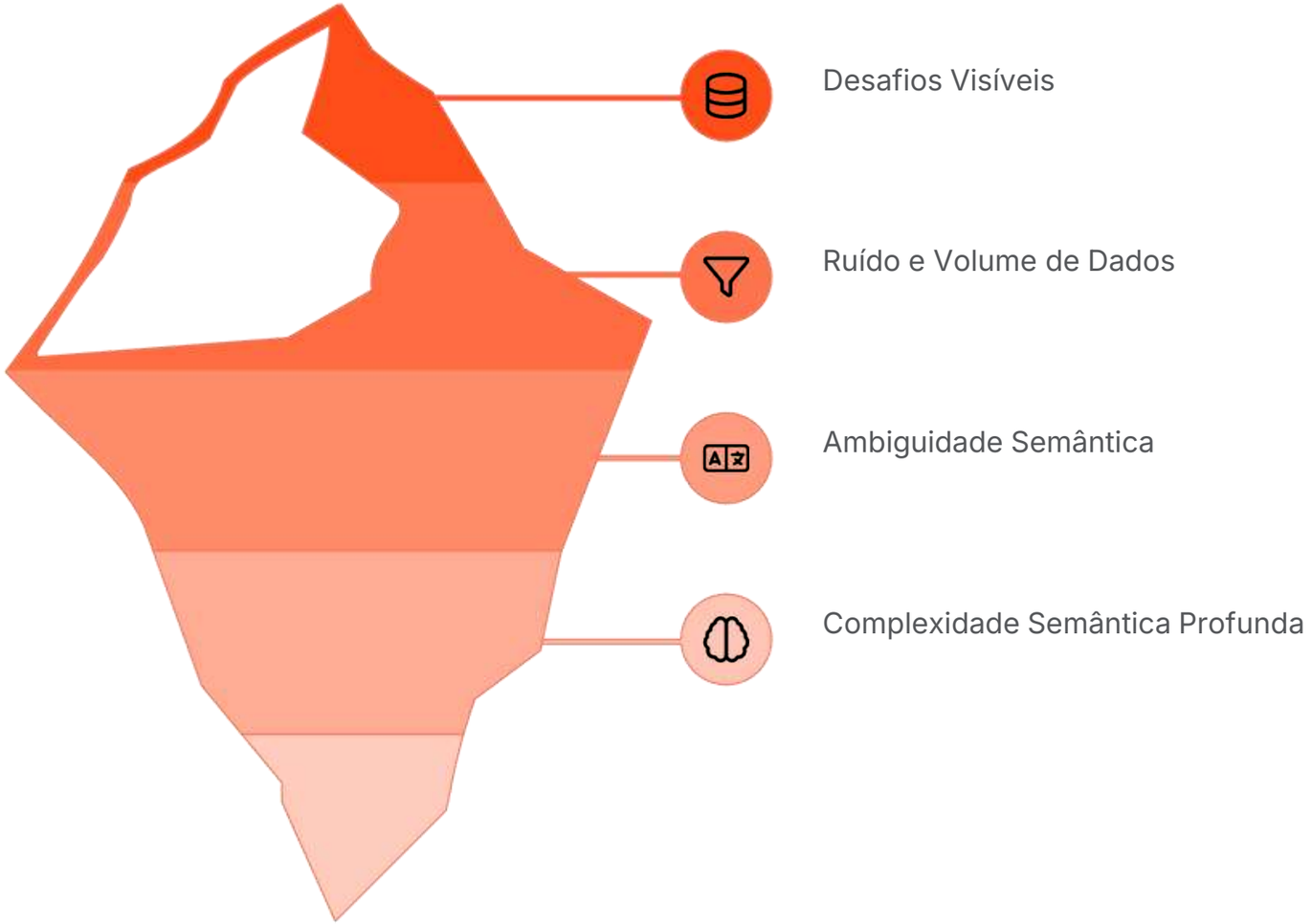
Cria resumos concisos de textos longos, como artigos científicos, notícias ou relatórios. Essencial para otimizar o consumo de informações e destacar pontos chave.

Detecção de Temas e Tendências

Identifica automaticamente os assuntos mais discutidos, mudanças na demanda do consumidor e o surgimento de novos interesses no mercado, orientando o desenvolvimento de produtos e estratégias.

Desafios em Text Mining

Embora poderoso, o Text Mining enfrenta obstáculos inerentes à complexidade da linguagem humana e ao volume de dados.



- Ambiguidade Linguística: Palavras com múltiplos significados (polissemia), dependendo do contexto.
- Ruído nos Dados: Textos com erros de digitação, abreviações informais, gírias e conteúdo irrelevante, exigindo pré-processamento robusto.

O Futuro do Text Mining: Rumo à Inteligência Conversacional

O Text Mining continua a evoluir, impulsionado por avanços em inteligência artificial e aprendizado de máquina, prometendo um futuro onde a interação homem-máquina será cada vez mais natural e eficiente.

A integração com IA e Machine Learning será ainda mais profunda, levando a uma automação avançada na extração e interpretação de texto. Veremos sistemas capazes de não apenas analisar, mas também gerar textos coerentes e contextualmente relevantes, como nos modelos de linguagem grandes (LLMs).

O foco se deslocará para a compreensão contextual e semântica profunda, permitindo que máquinas entendam nuances complexas da linguagem, como sarcasmo e ironia, com maior precisão. A personalização e a adaptabilidade em tempo real serão aprimoradas.

A aplicação de Text Mining se expandirá para novos domínios e idiomas, com a criação de modelos mais multilíngues e o suporte a dialetos e variações regionais. A ética e a privacidade dos dados textuais também serão temas centrais nas discussões futuras.

Ferramentas de Text Mining

AntWordProfiler

https://www.laurenceanthony.net/software/antwordprofiler/

Este recurso é uma ferramenta gratuita para analisar o nível de vocabulário e a complexidade de textos. O AntWord Profiler está disponível para download gratuito para Windows, Mac OS X ou Linux.

ConText

https://context.ischool.illinois.edu/

ConText é um aplicativo gratuito e de código aberto para executar uma variedade de técnicas de análise de texto, incluindo gráficos de rede e modelos de tópicos, com base em dados textuais.

https://gephi.org/

Gephi é uma plataforma aberta de visualização de grafos que suporta a exploração de todos os tipos de redes e sistemas complexos. O Gephi pode ser baixado gratuitamente em qualquer dispositivo Linux, Windows ou Mac OS X.

https://mimno.github.io/Mallet/index

Mallet é um pacote baseado em Java para processamento estatístico de linguagem natural, classificação de documentos, agrupamento, modelagem de tópicos, extração de informações e outras aplicações de aprendizado de máquina para texto.

Ferramentas de Text Mining

PhiloLogic3

https://sites.google.com/site/philologic3/

PhiloLogic é uma ferramenta de busca, recuperação e análise de texto completo desenvolvida pelo Projeto ARTFL e pelo Centro de Desenvolvimento de Bibliotecas Digitais (DLDC) da Universidade de Chicago. É um software gratuito que pode ser baixado para uma ampla gama de sistemas.

Textal

https://textal.org/

Textal é um aplicativo gratuito para smartphones que permite analisar sites, tweets e documentos para explorar a relação entre as palavras no texto por meio de uma interface intuitiva de nuvem de palavras. O aplicativo permite gerar gráficos e estatísticas, além de compartilhar dados e visualizações da maneira que você quiser. O Textal está disponível para download gratuito na App Store para seu dispositivo Apple iOS.

https://www.scrapy.org/

Scrapy é uma estrutura colaborativa e de código aberto para extrair os dados que você precisa de sites. Está disponível para download gratuito para Linux, Windows e Mac OS X.

TXM

https://sourceforge.net/projects/txm/

TXM é um ambiente de análise de texto/corpus, multiplataforma, gratuito, de código aberto e baseado em Unicode e XML, com cliente gráfico. Está disponível para download gratuito para Windows, Linux e Mac OS X. Possui uma gama abrangente de ferramentas de análise, como concordâncias, busca por colocação, lista de frequências, etc.