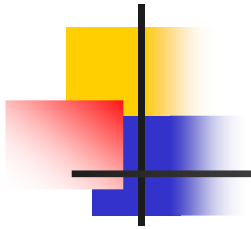
Ciência de Dados

KDD

Profa. Roseli A. F. Romero
– SCC-ICMC-USP

Prof. Dr. André C. P. L. F. de Carvalho
Dr. Isvani Frias-Blanco
ICMC-USP

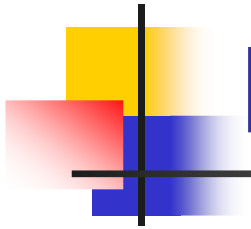


Tópicos do Módulo

- Introdução
- Descoberta de Conhecimento em Bases de Dados
- Etapas de KDD
- Mineração de Dados
- Aplicações

Introdução

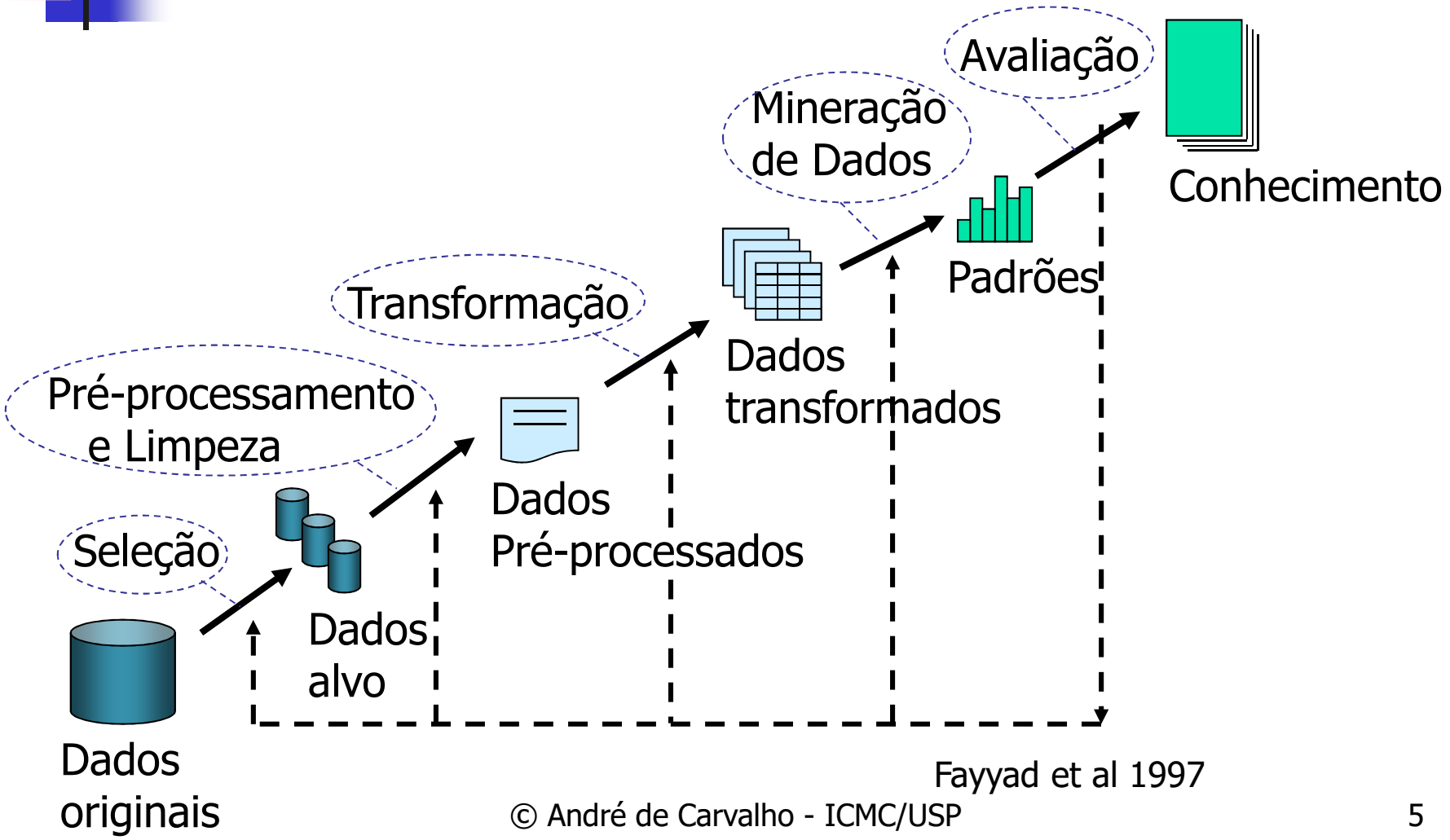
- Bases de Dados podem conter (esconder) dados preciosos
- Existe um interesse crescente em explorar esses dados armazenados
  - Descobrir conhecimento novo
  - Apoio à tomada de decisão



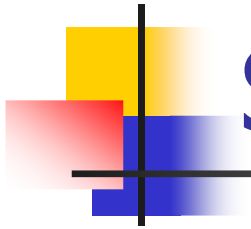
KDD

- *Knowledge Discovery in Databases*
- Processo de extrair conhecimento de dados
  - Útil
  - Novo
  - Válido
  - Potencialmente compreensível
- Processo interativo e iterativo
  - Várias etapas

KDD

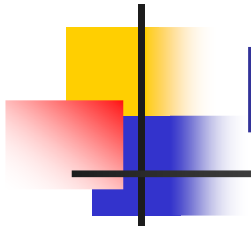


Fayyad et al 1997



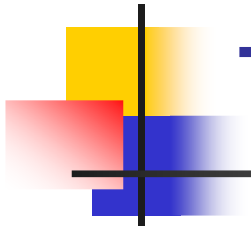
Seleção

- Extrai uma amostra do conjunto de dados para extração de conhecimento
  - Seleciona "manualmente" entre os dados disponíveis
    - Subconjunto de registros (instâncias ou exemplos)
    - Subconjunto de atributos considerados relevantes para o problema
      - Elimina atributos que sejam claramente irrelevantes



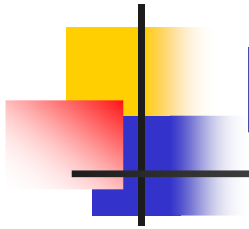
Pré-processamento e Limpeza

- Melhora a qualidade dos dados e facilita sua posterior utilização
- Engloba várias operações
  - Seleção "automática" de atributos
  - Conversão de valores
  - Tratamento de atributos com valores ausentes
  - Eliminação de dados duplicados
  - Detecção (e remoção) de ruído



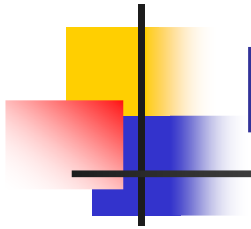
Transformação

- Inclui operações que modificam valores para um dado atributo
  - Cada operação deve ser aplicada a todos os valores do atributo
    - Em todos os objetos
  - Ex.: normalização, valor absoluto, ...



Mineração de Dados

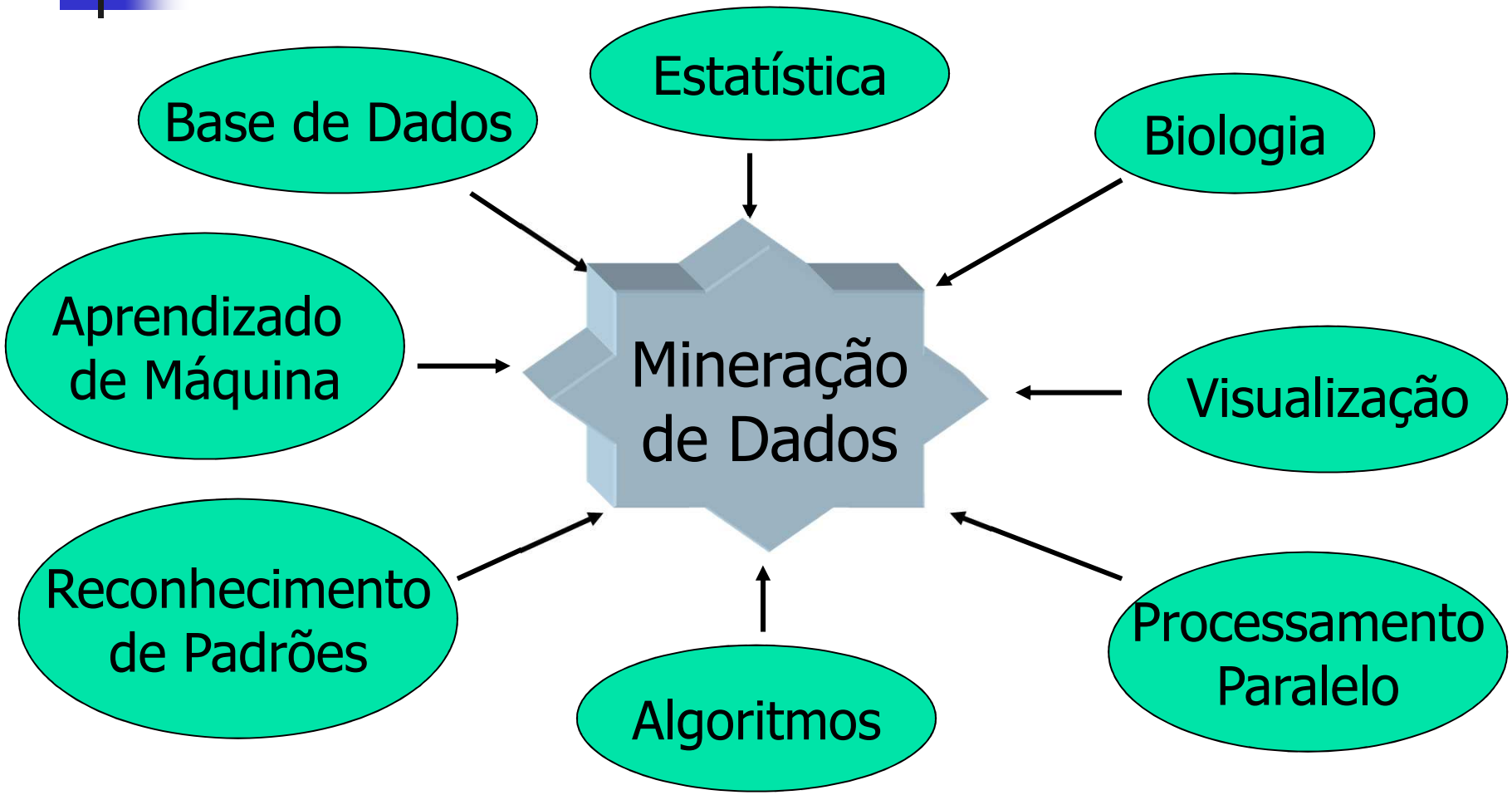
- Principal passo no processo de KDD
  - Mineração de Dados (DM) e KDD são frequentemente utilizados como sinônimos
- Difícil identificar fronteiras da etapa de MD no processo de KDD
  - Pré-processamento e transformação de dados são geralmente vistos como uma parte da MD

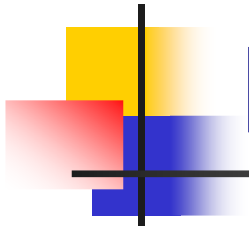


MD *X* KDD

- MD: ferramentas básicas utilizadas para extrair padrões de dados
- KDD: processo que engloba o uso dessas ferramentas, além de:
  - Seleção, pré-processamento, seleção, transformação dos dados
  - Interpretação e validação do conhecimento
    - Geração de conhecimento
    - Suporte à tomada de decisão

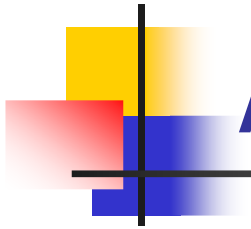
Mineração de Dados





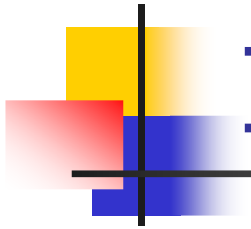
Mineração de Dados

- Outros termos utilizados para MD, KDD e CD
  - Extração de conhecimento
  - Descoberta de informação
  - Extração de padrões
  - Análise exploratória de dados
  - Analítica (*Data analytics* ou *analytics*)



Analítica

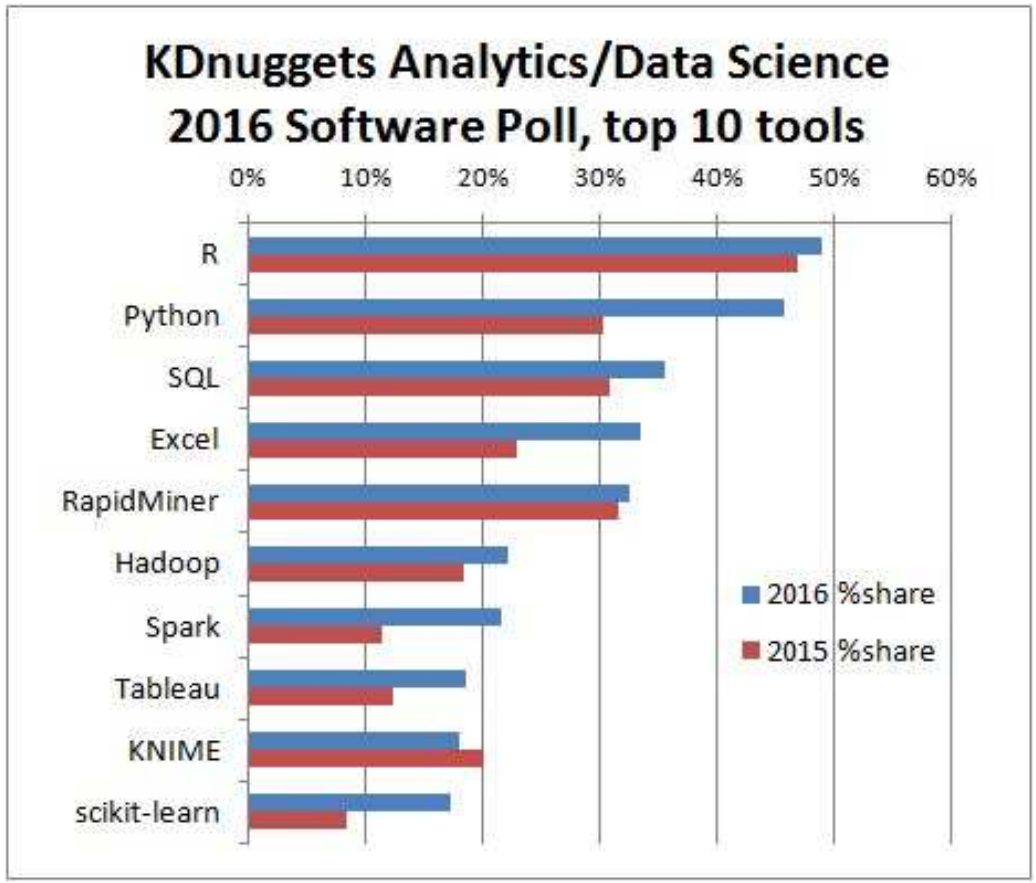
- Ciência que analisa dados crus para extrair padrões desses dados
  - Pode englobar coleta e organização dos dados
- Analítica preditiva (*predictive analytics*)
  - Extrai modelos (conhecimento) a partir de dados para realizar predições futuras
- Analítica descritiva (*descriptive analytics*)
  - Sumariza ou condensa dados para extrair conhecimento



Interpretação / Avaliação

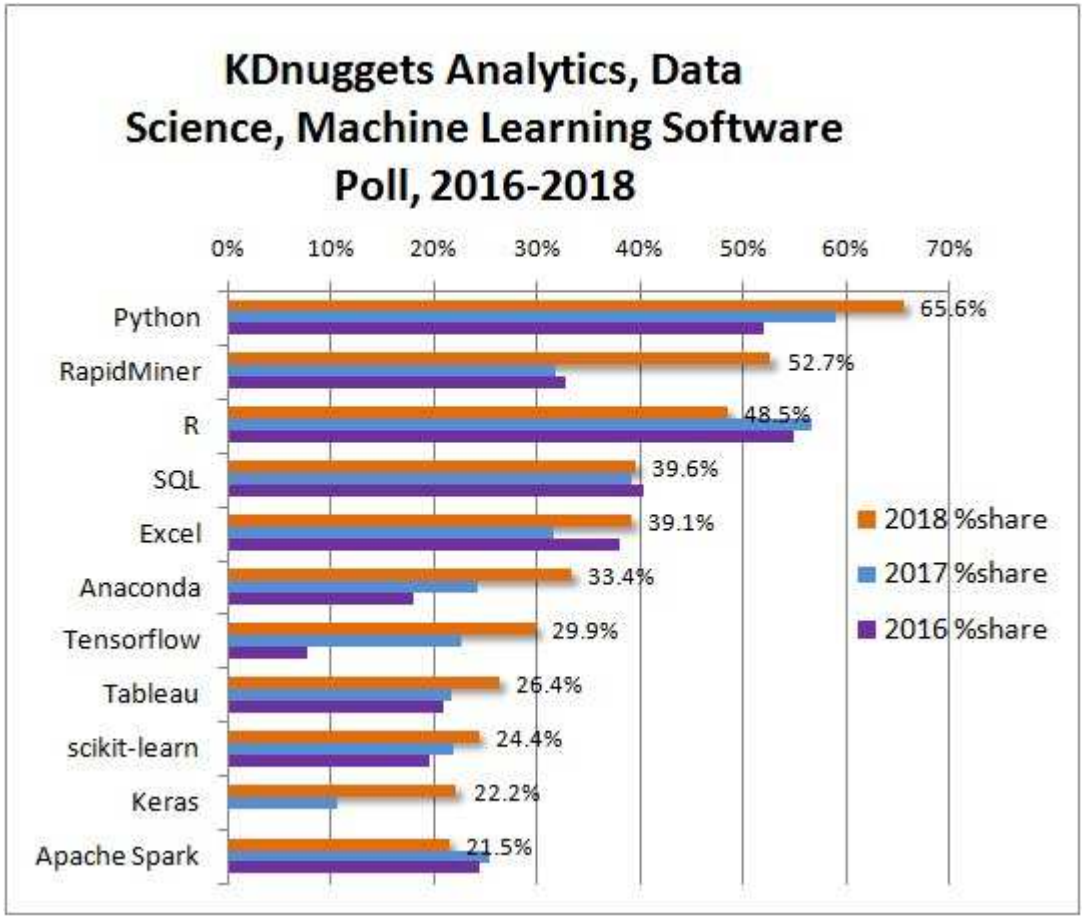
- Interpretação dos padrões encontrados na etapa de MD
  - Possível retorno a qualquer uma das etapas anteriores para iteração adicional
- Valida padrões encontrados
  - Importante consulta a um especialista
- Inclui análise estatística
- Ferramentas de visualização fornece um suporte importante

Ferramentas

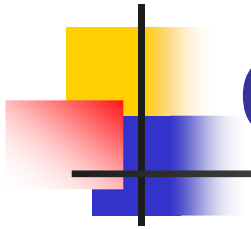


http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Ferramentas

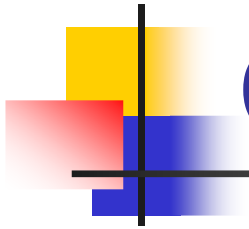


https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html



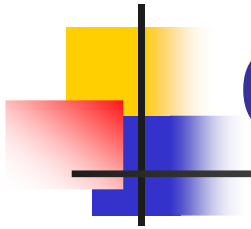
Custos em MD Preditivo

- 15% - coleta de dados
- 60% - limpeza e pré-processamento de dados
- 15% - construção e análise de modelos
- 5% - aplicação
- 5% - melhorias contínuas



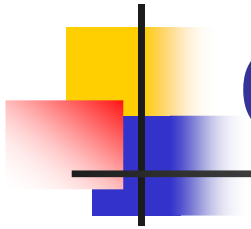
CRISP-DM

- Projeto CRISP-DM
  - *CRoss-Industry Standard Process for Data Mining*
  - Concebido em 1996 por:
    - Daimer-Chrysler
      - Aplicava MD em suas operações de negócios
    - SPSS
      - Prestava serviço de MD desde 1990
      - Desenvolveu primeira ferramenta comercial de MD (*Clemetine*)
    - NDR
      - Tinha o propósito de adicionar valor a sua enorme BD



CRISP-DM

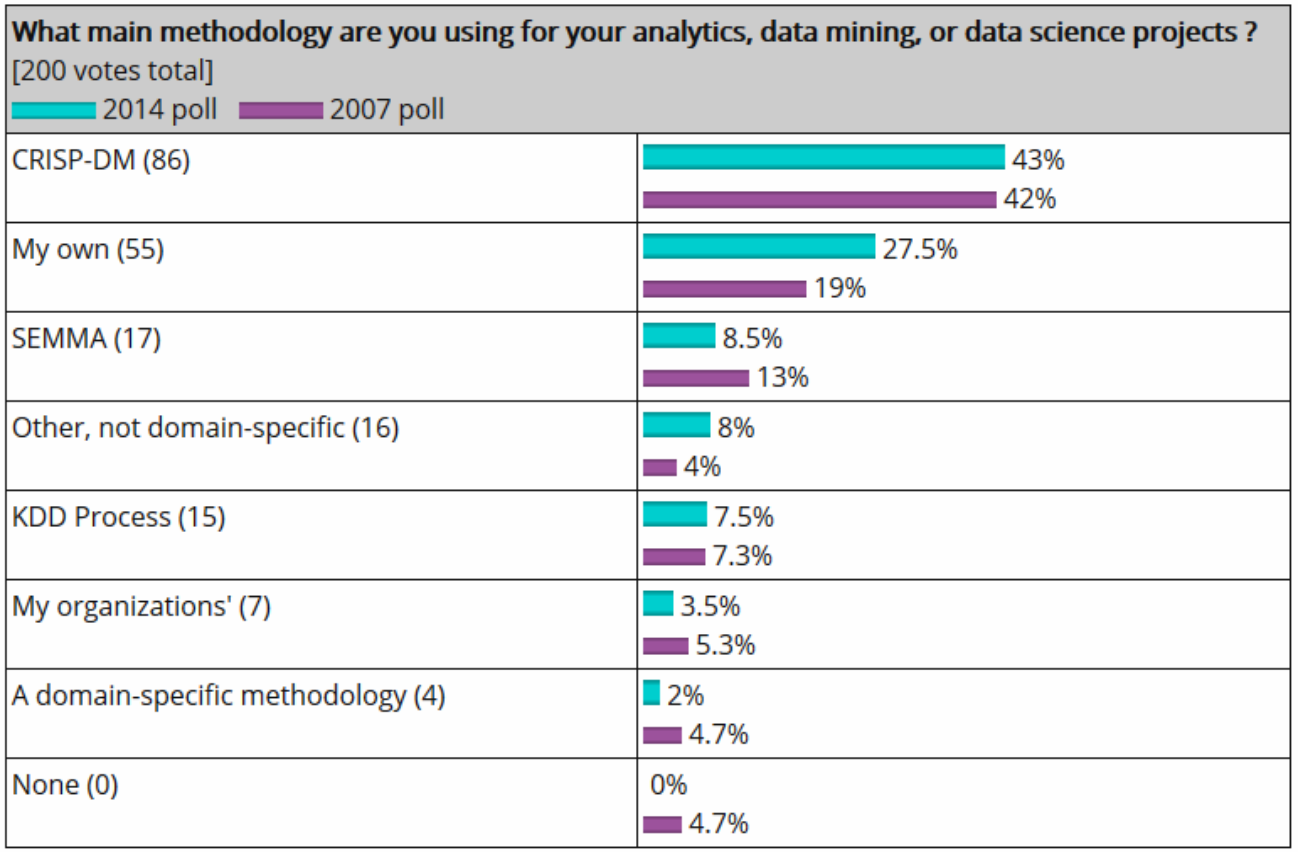
- Projeto CRISP-DM
  - Desenvolveu um novo fluxo de processos para descoberta de conhecimento
    - A partir do processo anterior (KDD)
      - Fayyad, Piatesky-Shapiro and Smyth
    - Em resposta a requisitos de usuários
    - Definiu e validou processo de MD utilizado em vários setores industriais

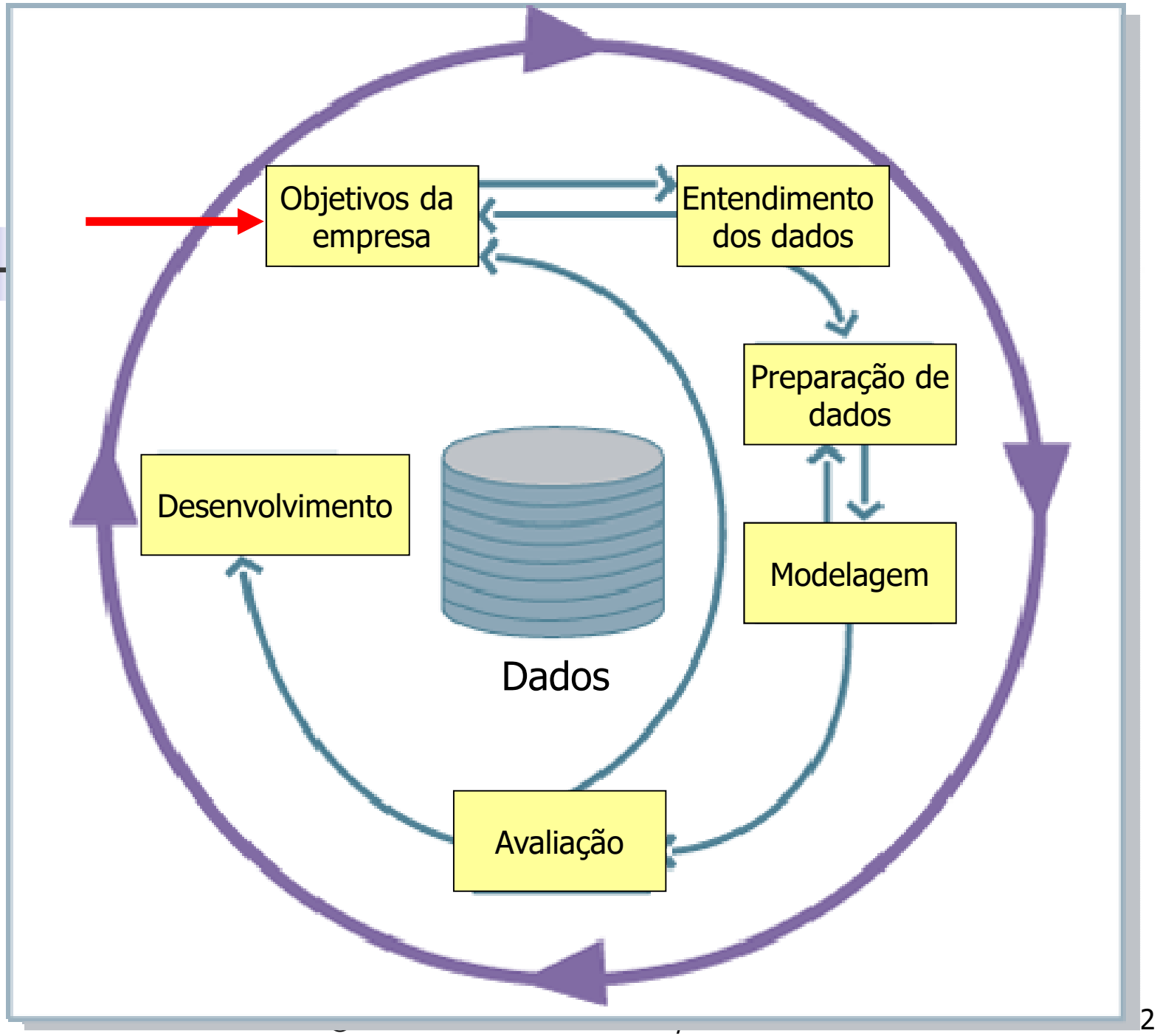


CRISP-DM

- Metodologia procura tornar os projetos
  - Mais rápidos
  - Mais baratos
  - Mais confiáveis
  - Mais facilmente gerenciáveis
- Pode ser aplicada a pequenos projetos
- Metodologia mais popular em MD
  - Metodologia padrão da indústria

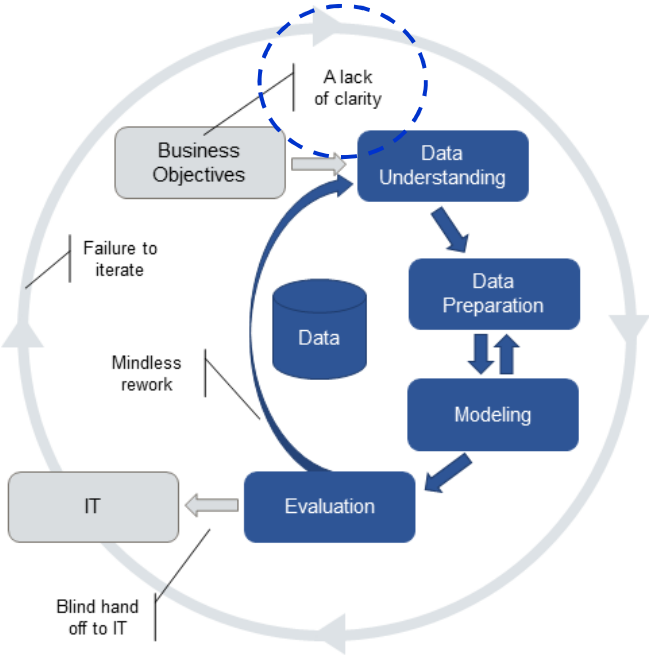
Metodologias mais populares




Objetivos da empresa
Entendimento dos dados
Preparação de dados
Modelagem
Avaliação
Desenvolvimento
Dados

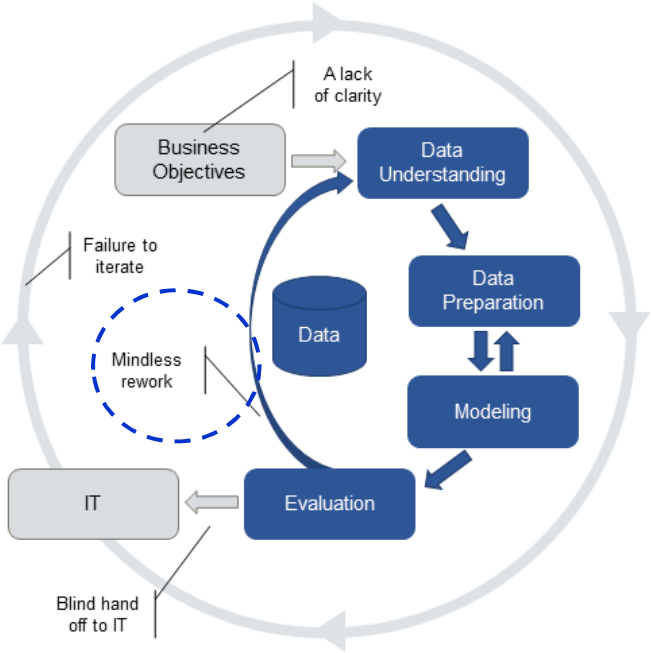
Problemas na aplicação de CRISP-DM

- Entender objetivos da empresa (aplicação)
  - Pode confundir mais que ajudar
  - Ao invés disso, equipe de CD deveria entender em detalhes os problemas da empresa
    - e como CD pode ajudar
    - Senão, pode gerar modelos interessantes, mas que não atacam os problemas



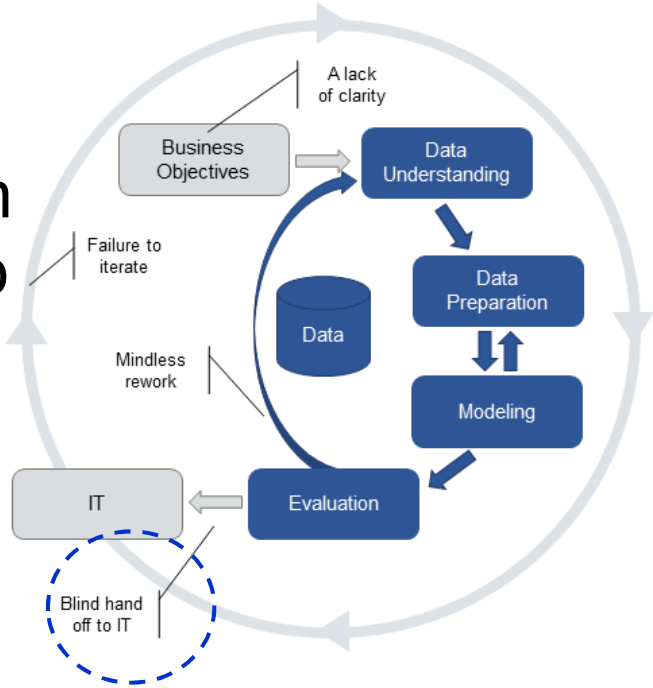
Problemas na aplicação de CRISP-DM

- Avaliação na direção errada
  - Retrabalho sem necessidade
    - Algumas equipes apenas olham o desempenho preditivo dos modelos
      - Se o modelo tem bom desempenho preditivo, ele deve ser bom
        - Depois vê que não é verdade e busca relação entre resultados e objetivos da empresa
          - Se não casam:
            - Busca novos dados e tenta outras técnicas de CD
            - Ao invés de avaliar como modelo ataca os problemas



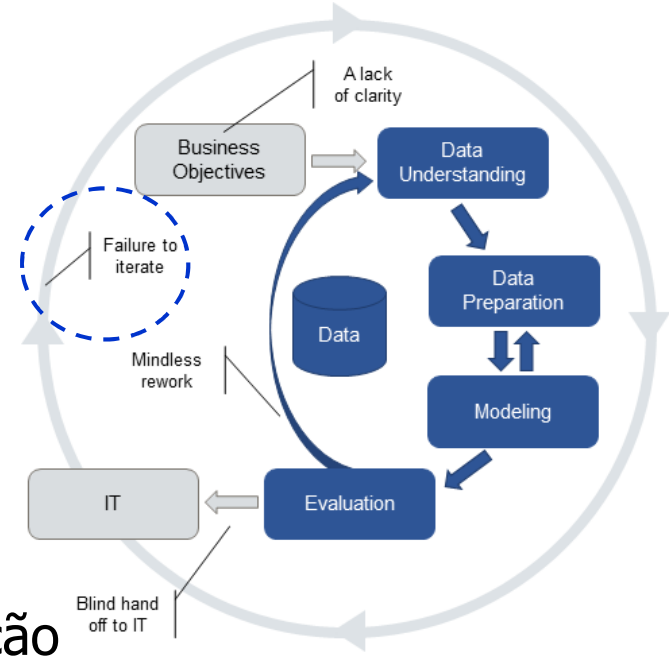
Problemas na aplicação de CRISP-DM

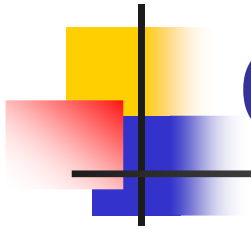
- Ignora implementação e uso
  - Equipe envia modelo para TI sem preocupação com implementação e aplicação a dados reais
    - Equipe não conversa antes com TI
      - *"Implementação e uso é problema dos outros (TI)"*
      - Aumenta custo (tempo e dinheiro)
      - Reduz chance de modelos beneficiarem a empresa



Problemas na aplicação de CRISP-DM

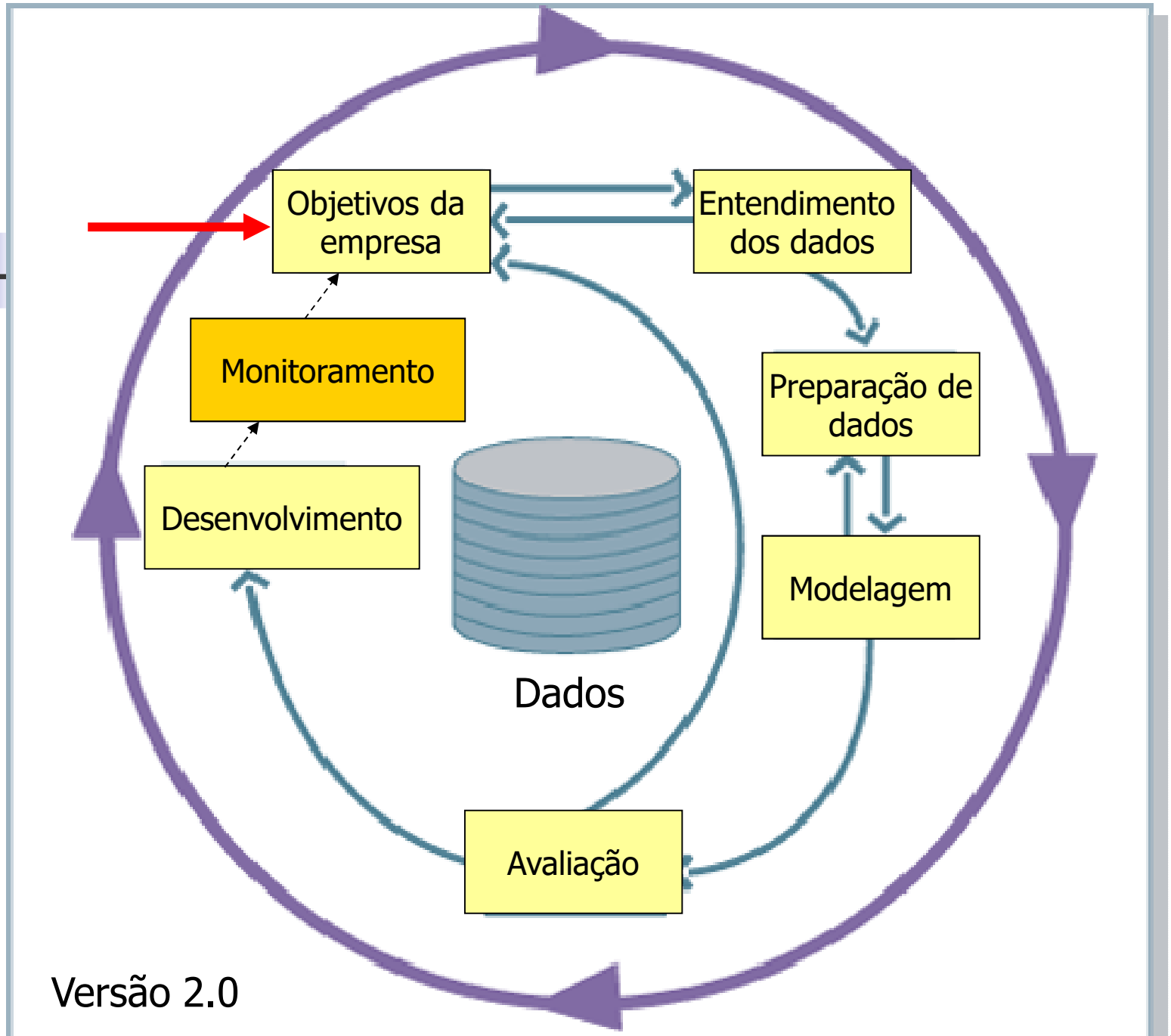
- Sem iteração, sem atualização
  - Modelos podem se tornar obsoletos
    - Devido a mudanças que podem ocorrer ao longo do tempo
    - Modelos precisam ser atualizados para manter seu valor
    - Equipe de CD fica adiando atualização
      - Não monitora nem atualiza o modelo, reduzindo contribuição da CD
      - Constrói solução que parece boa no início, mas que, no final, não beneficia a empresa

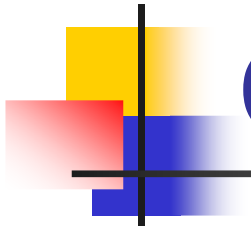




CRISP-DM 2.0

- SIG (*special interest group*) formado entre 2006 e 2008
- Mudanças estudadas
  - Divisão da fase de preparação de dados
  - Métodos de avaliação dentro da fase de modelagem
  - Fase de avaliação associada a avaliação na empresa
  - Inclusão de fase de monitoramento

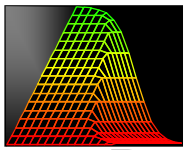

Objetivos da empresa
Entendimento dos dados
Preparação de dados
Modelagem
Avaliação
Desenvolvimento
Monitoramento
Dados
Versão 2.0



CRISP-DM 2.0

- Não foi lançado até 2005
- SIG foi desfeito
  - Website CRISP-DM.org não esta mais ativo
- IBM criou nova metodologia
  - Refina e estende CRISP-DM
  - *Analytics Solutions Unified Method for Data Mining/Predictive Analytics* (ASUM-DM)

Produtos para MD



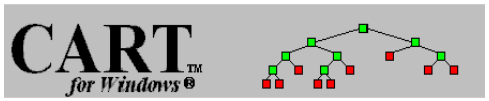
PREDICTIVE DYNAMIX

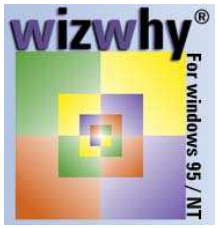
QuickTime™ and a
GIF decompressor
are needed to see this picture.

QuickTime™ and a
GIF decompressor
are needed to see this picture.

Model 1

KnowledgeMiner 5.0





PolyAnalyst 4.5

NeuroShell®2

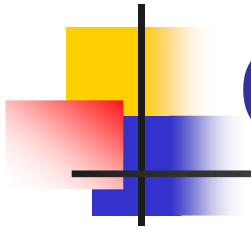
QuickTime™ and a
GIF decompressor
are needed to see this picture.

Mais Produtos

PRW

DBMiner Insight

Partek Pro 5.0



Considerações Finais

- Expansão do volume de dados armazenados
  - Necessidade de extrair conhecimento dos dados
  - KDD, CRISP-DM, ...
  - Cuidado com promessas exageradas
- Leitura
  - Knowledge Discovery and Data Mining: Towards a Unifying Framework, U. Fayyad, P. Smyth, and G. Piatetsky-Shapiro, .2nd International Conference on Knowledge Discovery and Data Mining, 1996
  - Integrating and Updating Domain Knowledge with Data Mining, Carsten Pohle, 2003.

Perguntas