SCC0173- Mineração de Dados

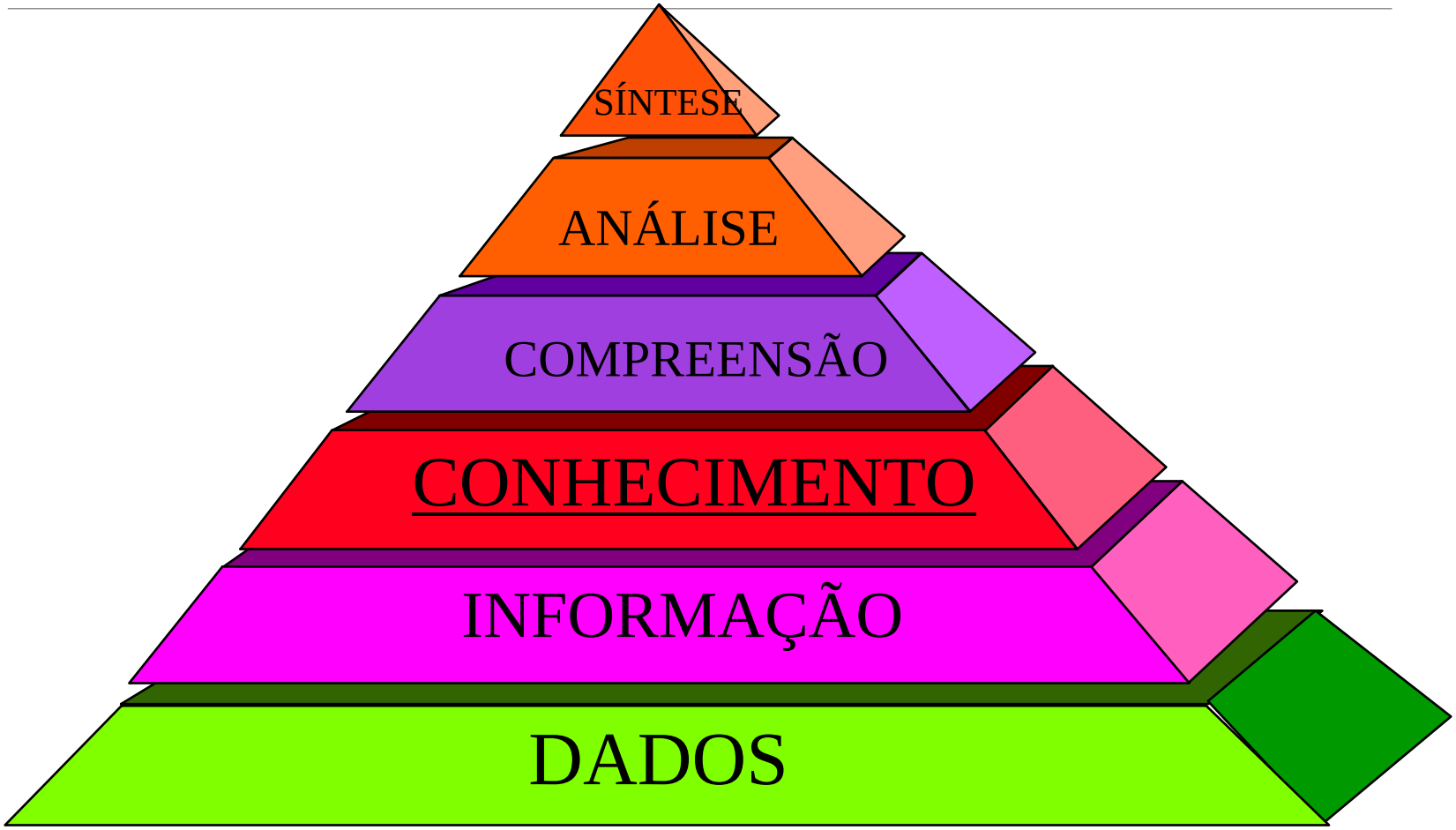
KDD e MD

Docente: Solange Rezende
PAE: Brucce

Introdução

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

De Dados à manipulação de Conhecimento…



O que é um dado?

- Dado
  - Exemplo, objeto, registro
  - Estrutura fundamental sobre a qual um sistema de informação é construído
  - Não tem um significado associado a ele
    - Ex.: 345,43

Exemplo – Carros

muito alto,muito alto,2,2,pequeno,baixo,nenhum
muito alto,alto,3,5 ou mais,grande,baixo,nenhum
muito alto,baixo,3,4,grande,baixo,nenhum
médio,baixo,4,2,pequeno,alto,nenhum
médio,baixo,3,4,pequeno,médio,médio
alto,alto,2,4,grande,médio,médio
baixo,baixo,5 ou mais,4,pequeno,médio,médio
baixo,médio,4,4,pequeno,médio,médio
baixo,médio,4,4,grande,médio,bom
baixo,baixo,4,5 ou mais,grande,médio,bom
médio,baixo,2,4,pequeno,alto,bom
baixo,médio,4,4,grande,alto,muito bom
médio,médio,2,4,grande,alto,muito bom
baixo,baixo,5 ou mais,5 ou mais,grande,alto,muito bom

Informação

- Dados com um significado associado
  - Tornam os dados úteis em uma tomada de decisão
  - A transformação de dados em informação
- Geralmente pela apresentação dos dados em uma forma compreensível para o usuário
  - Informação é criada quando é associado um significado a um conjunto de dados

Exemplo - Carros

- Preço
  - Compra: v-alto, alto, médio, baixo
  - Manutenção: v-alto, alto, médio, baixo
- Características técnicas
  - Conforto
    - # portas: 2, 3, 4, 5-5 ou mais
    - # pessoas: 2, 4, 5 ou mais
    - Espaço do porta malas: pequeno, médio, grande
  - Segurança: baixo, médio, alto
- Aval. do carro: nenhum, médio, bom, muito bom

Exemplo - Carros

muito alto,muito alto,2,2,pequeno,baixo,nenhum
muito alto,alto,3,5 ou mais,grande,baixo,nenhum
muito alto,baixo,3,4,grande,baixo,nenhum
médio,baixo,4,2,pequeno,alto,nenhum
médio,baixo,3,4,pequeno,médio,médio
alto,alto,2,4,grande,médio,médio
baixo,baixo,5 ou mais,4,pequeno,médio,médio
baixo,médio,4,4,pequeno,médio,médio
baixo,médio,4,4,grande,médio,bom
baixo,baixo,4,5 ou mais,grande,médio,bom
médio,baixo,2,4,pequeno,alto,bom
baixo,médio,4,4,grande,alto,muito bom
médio,médio,2,4,grande,alto,muito bom
baixo,baixo,5 ou mais,5 ou mais,grande,alto,muito bom

- Preço
  - Compra: v-alto, alto, médio, baixo
  - Manutenção: v-alto, alto, médio, baixo
- Características técnicas
  - Conforto
    - # portas: 2, 3, 4, 5-5 ou mais
    - # pessoas: 2, 4, 5 ou mais
    - Espaço porta malas: pequeno, médio, grande
  - Segurança: baixo, médio, alto
- Aval. do carro: nenhum, médio, bom, muito bom

Exemplo - Promotores

+,S10,tactagcaatacgcttgcgttcggtggttaagtatgtataatgcgcgggcttgtcgt
+,AMPC, tgctatcctgacagttgtcacgctgattggtgtcgttacaatctaacgcatcgccaa
+,AROH, gtactagagaactagtgcattagcttatttttttgttatcatgctaggcggcg
+,DEOP2,aattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaatactaacaaactc
+,LEU1_TRNA,tcgataattaactattgacgaaaagctgaaagactagaatgcgcctccgtggtag
+,MALEFG,aggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgtttaggt
-, 296,aggcatgtaaacgtcttcgtagcgcatcagtgctttcttactgtgagtacgcgag
-, 648,ccgagtaggcttagagagcatgtcagcctcgacaacttgcataaatgctttcttg
-, 230,cgctaggactttcttgttgattttccatgcggtgttttgcgcaatgttaatcgcttt
-,1163,tatgggaacgagtcaatcagggctttgactctggtattactgtgaacattatt
-,1321,agagggtgtactccaagaagaggaagatgaggctagacgtctctgcatggagtatga
-, 663,gagagcatgtcagcctcgacaacttgcataaatgctttcttgtagacgtgccctacg

Análise de dados

- Análise dos dados por seres humanos
  - Falta de especialistas
  - Custo elevado
  - Subjetividade
  - Volume
- Técnicas tradicionais para análise
  - Sistemas de gerenciamento de bancos de dados
  - Planilhas

Análise de dados

- Técnicas tradicionais de análise de dados permitem apenas consultas simples
  - Quantos itens de um produto em particular foram vendidos em um dado dia?
  - Não conseguem responder consultas do tipo:
    - Dadas características de um carro, devo comprá-lo?
    - Que tecidos podem estar com tumor?
    - Qual a estrutura terciária de uma nova proteína
- Técnicas mais sofisticadas, capazes de extrair conhecimento de grandes conjuntos de dados

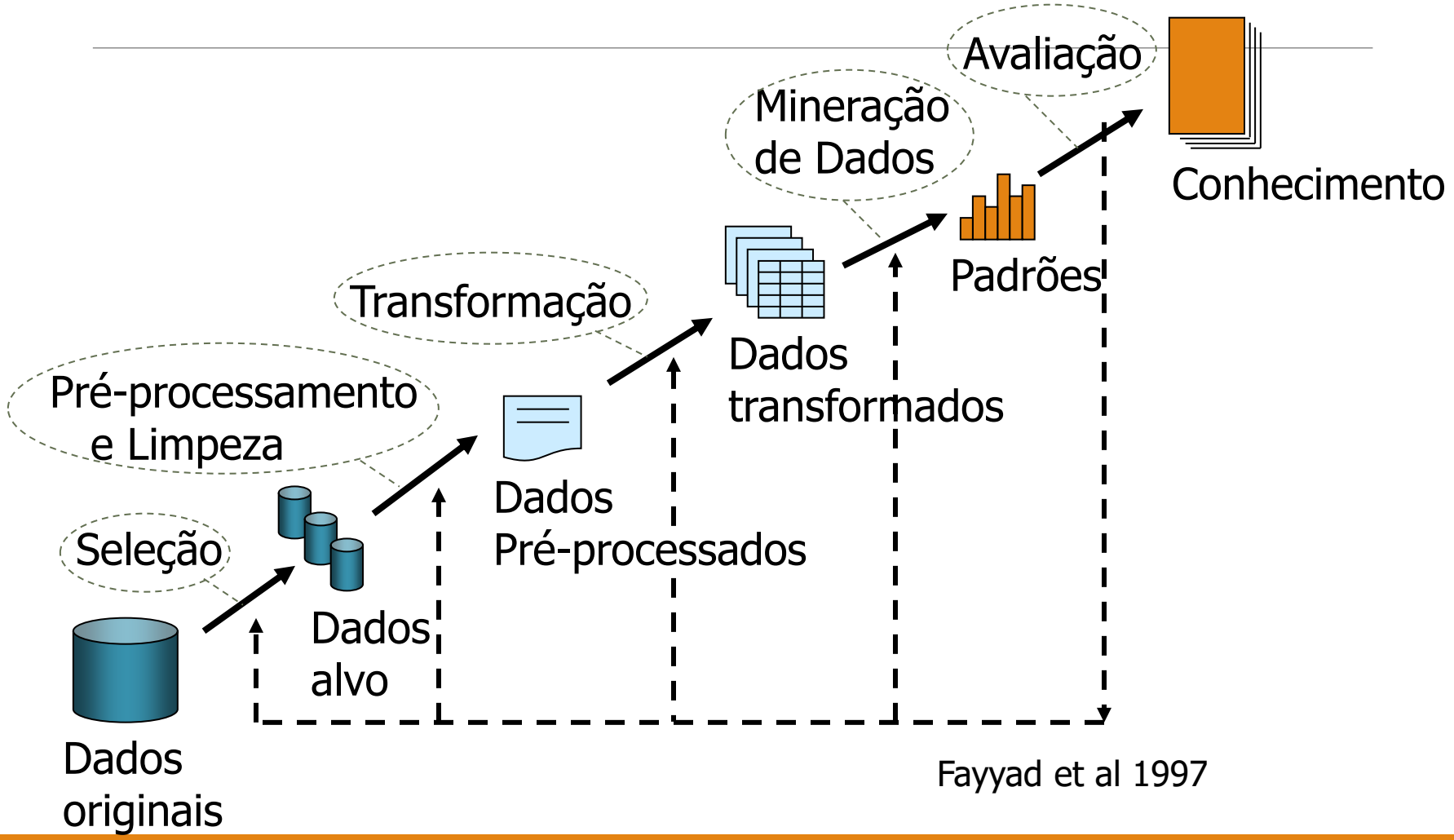
KDD

- Descoberta de conhecimento em Bancos de dados (BDs)
  - *Knowledge Discovery in Databases*
- Área de pesquisa em expansão
- Teorias e ferramentas computacionais para extrair informação útil de BDs
  - Informação útil = conhecimento

KDD

- Processo de encontrar em dados padrões
  - Úteis
  - Novos
  - Válidos
  - Potencialmente compreensíveis
- Processo interativo e iterativo
  - Várias etapas
    - Uma delas é Mineração de Dados

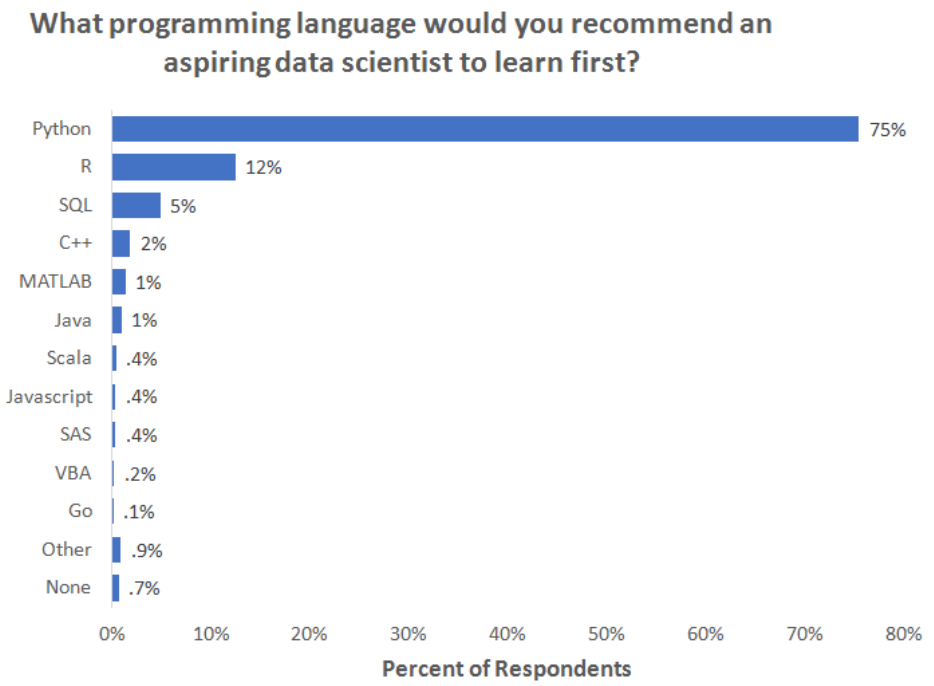
KDD


Avaliação
Mineração de Dados
Conhecimento
Padrões
Transformação
Dados transformados
Pré-processamento e Limpeza
Dados Pré-processados
Seleção
Dados alvo
Dados originais
Fayyad et al 1997

Ferramentas

http://businessoverbroadway.com/wp-content/uploads/2019/01/programming_languages_recommended.png

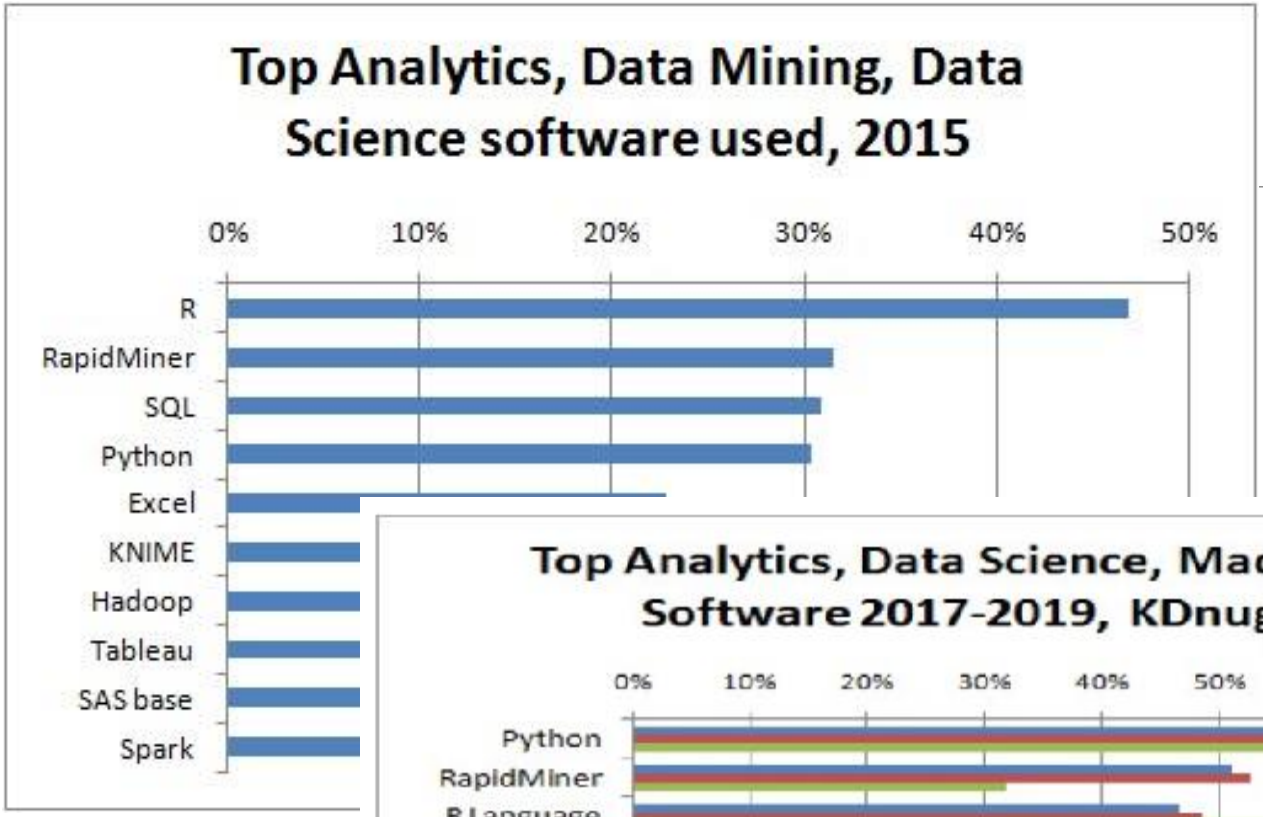
Your primary programming language for Analytics, Data Mining, Data Science tasks: [512 voters]	
2015 primary programming language:	
R (and its packages) (263)	51% (of 2015 votes)
Python (including scikit-learn and other libraries) (151)	29%
Other (Java, MATLAB, SAS, Scala, etc) (89)	17%
none (9)	1.8%
2014 primary programming language:	
R (and its packages) (237)	2014 votes)
Python (including scikit-learn and other libraries) (117)	23%
Other (Java, MATLAB, SAS, Scala, etc) (118)	23%
none (40)	7.8%



Note: Data are from the 2018 Kaggle ML and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018.
A total of 23859 respondents completed the survey; the percentages in the graph are based on a total of 18788 respondents who provided an answer to this question.

Ferramentas

https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html



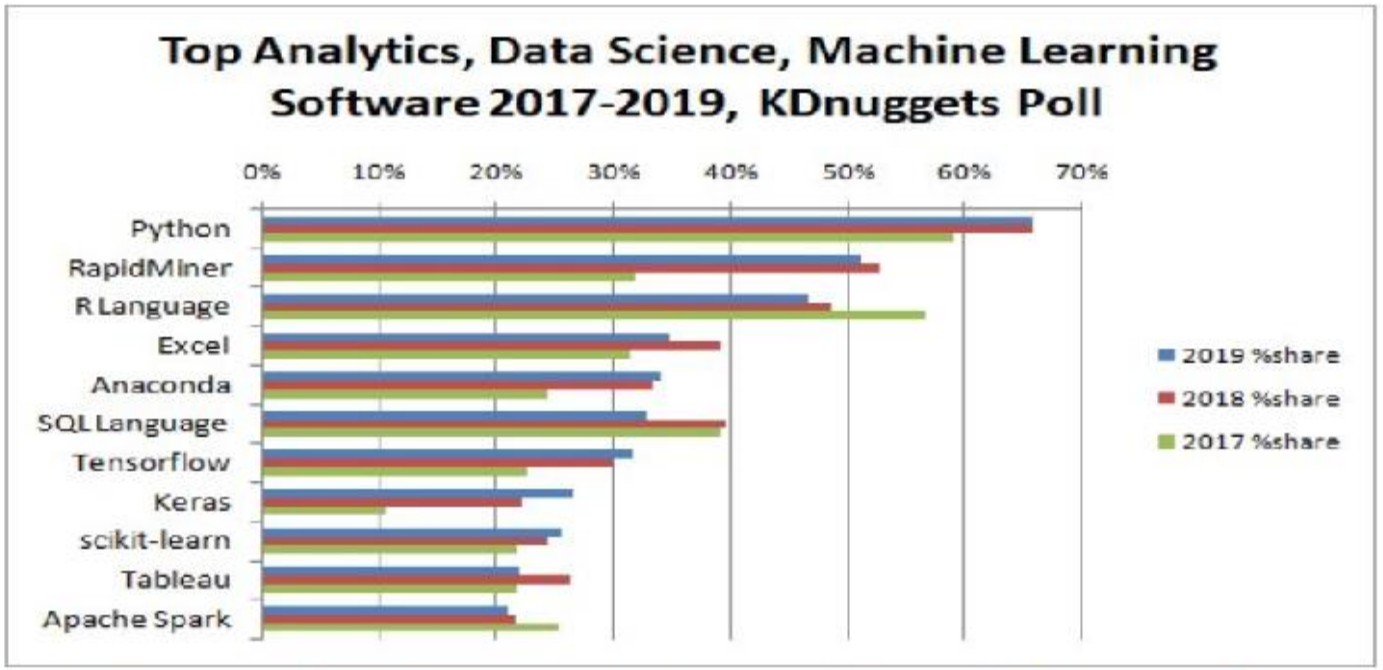


Fig 1: KDnuggets Analytics/Data Science 2019 Software Poll: top tools in 2019, and their share in the 2017, 2018 polls

Investimentos em MD Preditivo

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

CRISP-DM

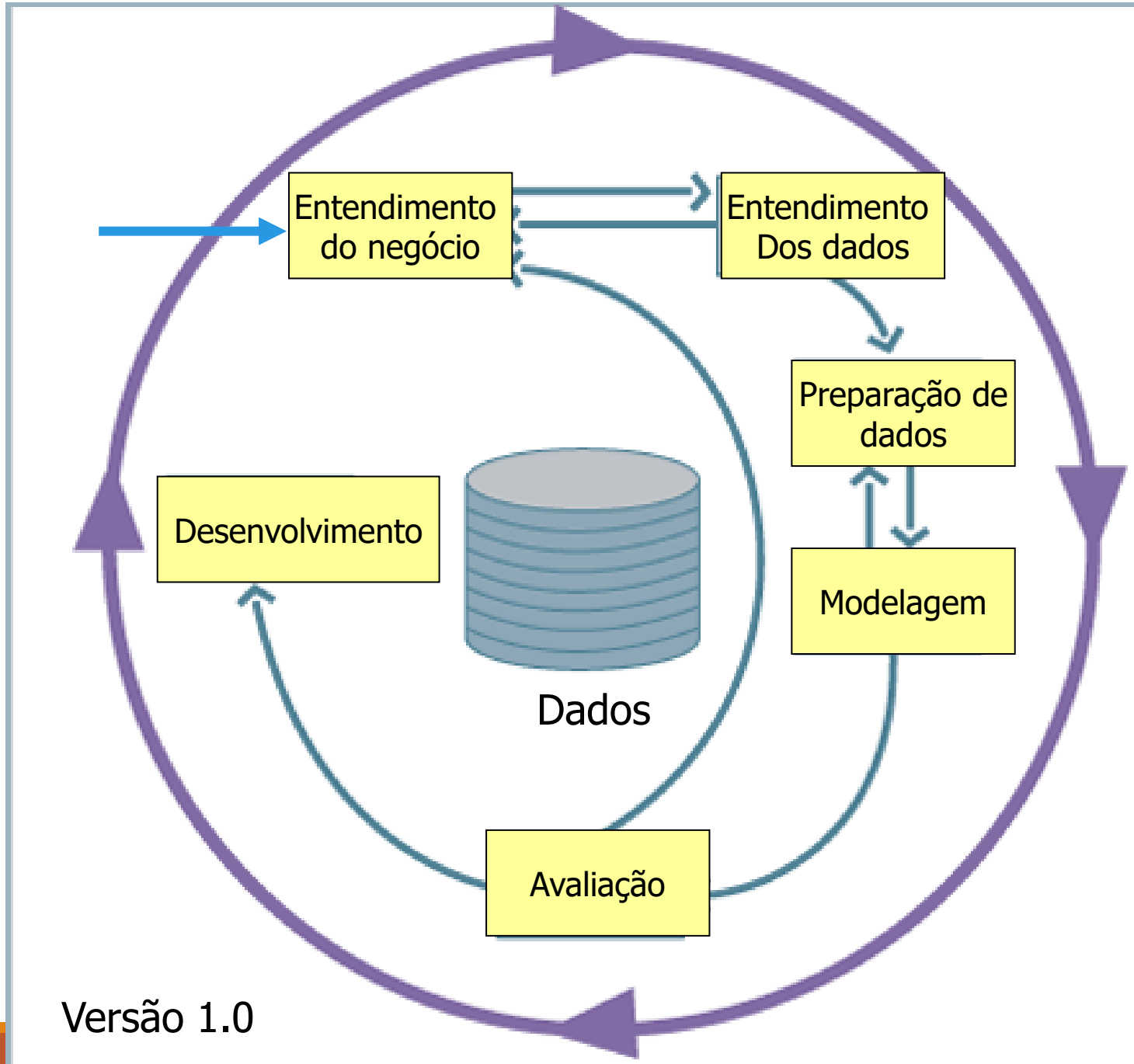
- Projeto CRISP-DM
  - *CRoss-Industry Standard Process for Data Mining*
  - Concebido em 1996 por:
    - Daimer-Chrysler
      - Aplicava MD em suas operações de negócios
    - SPSS
      - Prestava serviço de MD desde 1990
      - Desenvolveu primeira ferramenta comercial de MD (*Clemetine*)
    -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

CRISP-DM

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

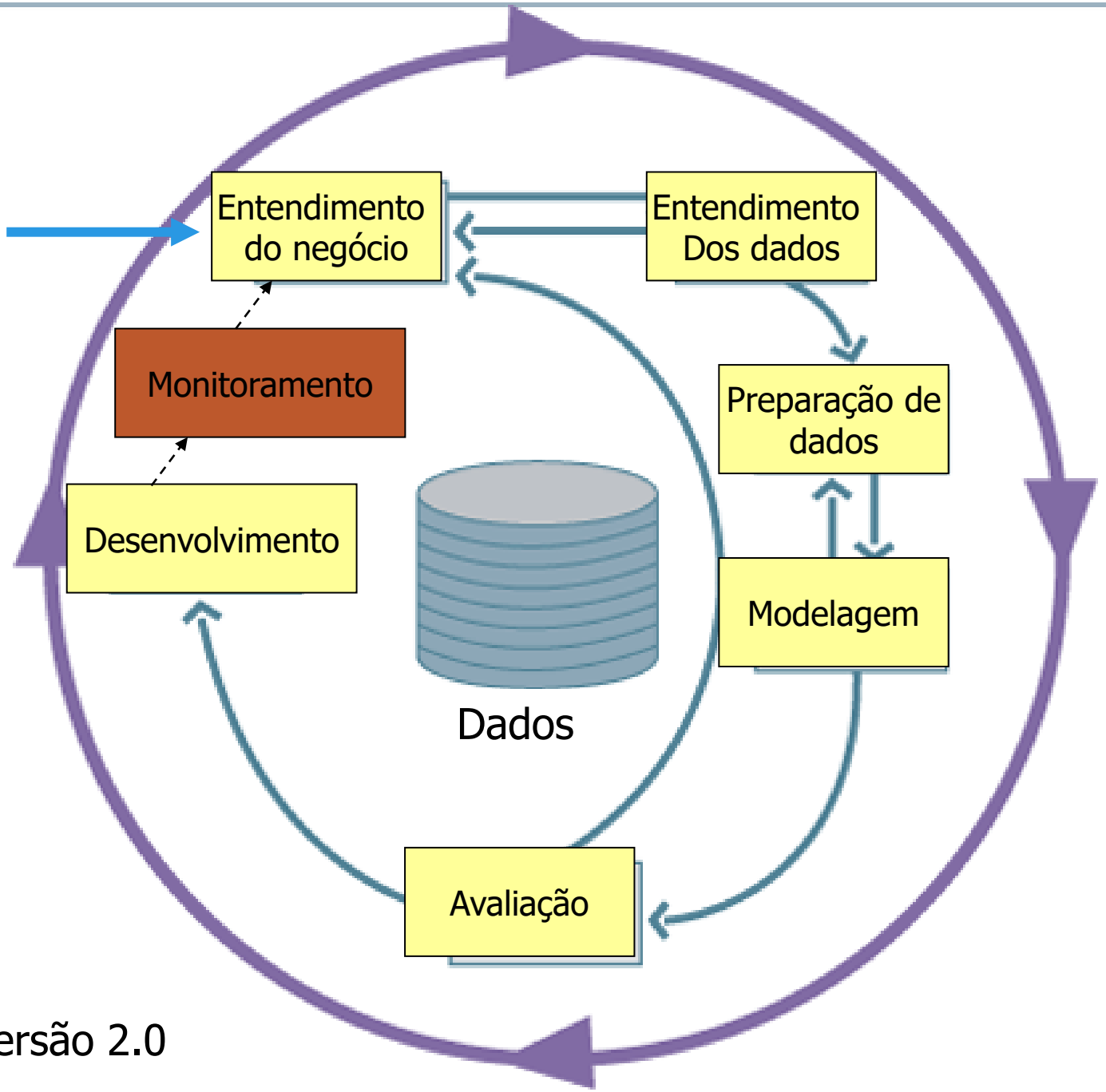
CRISP-DM

- Nova metodologia procura tornar os projetos
  - Mais rápidos
  - Mais baratos
  - Mais confiáveis
  - Mais facilmente gerenciáveis
- Pode ser aplicada a pequenos projetos
- Metodologia padrão da indústria


Entendimento do negócio
Entendimento Dos dados
Preparação de dados
Modelagem
Avaliação
Desenvolvimento
Dados
Versão 1.0

CRISP-DM 2.0

- Mudanças estudadas
  - Divisão da fase de preparação de dados
  - Métodos de avaliação dentro da fase de modelagem
  - Fase de avaliação será associada a avaliação na empresa
    - Centro de pesquisa, laboratório, hospital,...
  - Inclusão de **fase de monitoramento**

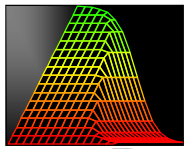

Entendimento do negócio
Entendimento Dos dados
Monitoramento
Preparação de dados
Desenvolvimento
Modelagem
Dados
Avaliação

Versão 2.0

Mineração de Dados na prática

Corresponde ao **Entendimento do Negócio** e **Entendimento dos Dados**

Produtos de MD

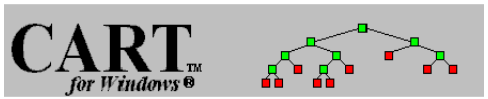


PREDICTIVE DYNAMIX

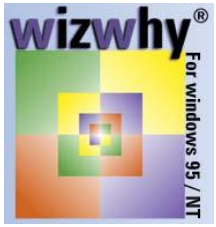
QuickTime™ and a
GIF decompressor
are needed to see this picture.

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

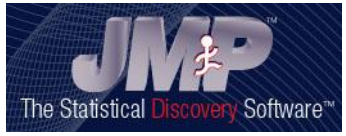
Model 1



KnowledgeMiner 5.0



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



Mais Produtos

DataEngine

DataMIND

pcOLPARS

Urban Science
GainSMARTS

PRW

ModelQuest Enterprise™

TextSense

Intelligent Miner for data

comotion

DBMiner Insight

NETEZZA

Partek Pro 5.0

Considerações Finais

- Expansão do volume de dados armazenados
- Necessidade de extrair conhecimento dos dados
- KDD é cada vez mais usado
- Cuidado com promessas exageradas
- Leitura – Fonte definição de KDD
  - 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

Perguntas