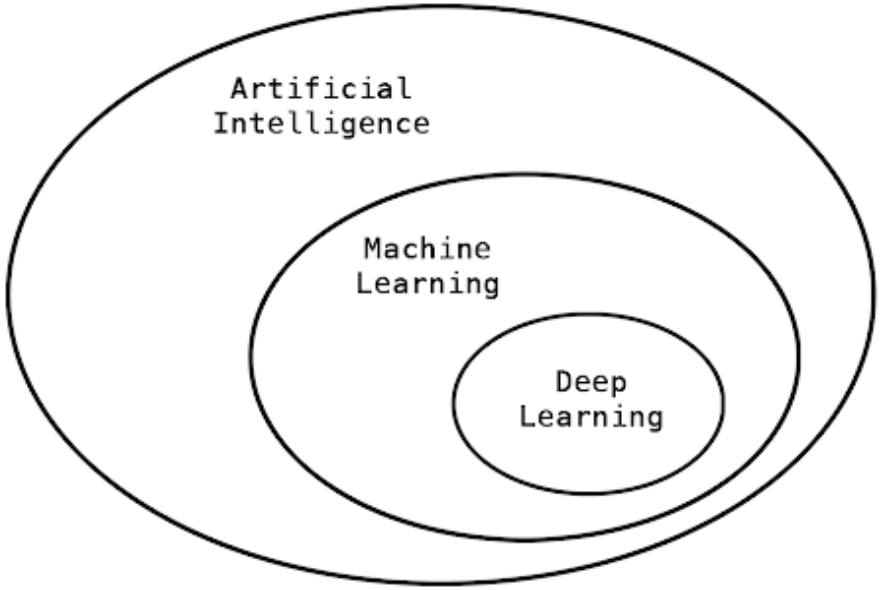
Introdução à Ciência de Dados

Fátima Rodrigues (DEI/ISEP)
mfc@isep.ipp.pt

Inteligência Artificial

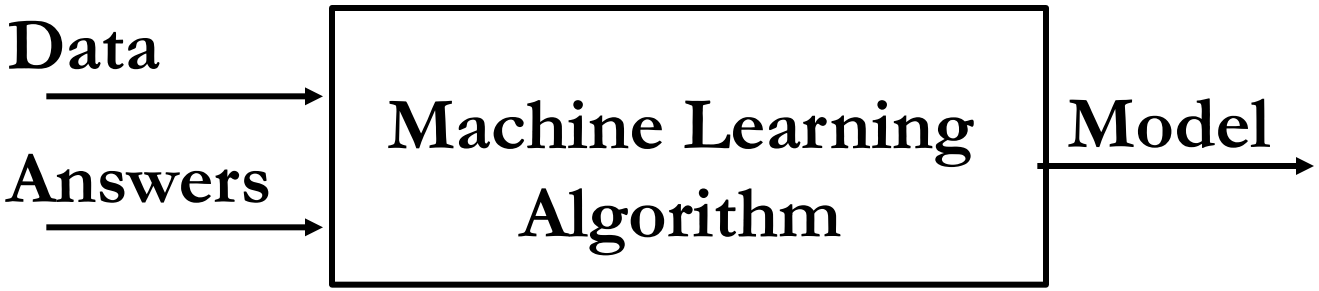


- **Inteligência Artificial** é a ciência que treina máquinas para realizar tarefas humanas

- **Machine Learning** foca-se na criação de algoritmos que têm a habilidade de aprender com os dados sem serem explicitamente programados

Machine Learning/Aprendizagem Máquina

Novo paradigma de programação



Os algoritmos de Machine Learning aprendem a partir de dados

Evolução da Tecnologia de Bases de Dados

- **1950s**: Primeiros computadores, usados no censos da população Americana
- **1960s**: Coleção de dados, criação de bases de dados (modelo hierárquico e de rede)
- **1970s**: Modelo Relacional, Implementação DBMS
- **1980s**: RDBMS, modelos de dados avançados (relacional estendido, OO, dedutivo, etc.) e DBMS orientados a aplicações (espacial, científica, etc.)
- **1990s**: armazéns de dados (OLAP), bases de dados multimédia, bases de dados WEB (surge Data Mining)
- **2000s:** século da informação
  - Fluxo constante de dados: sensor networks, web logs, mobile SMS, computer network traffic
  - WEB 2.0 (mensagens e logs gerados de blogs, redes sociais, jornais on-line)

Dados Armazenados

- Transações de negócio suportadas por cartões magnéticos
- Dados científicos
- Dados pessoais e médicos
- Vídeo, áudio, imagens
- Dados recolhidos via satélite espaciais, sensores remotos
- Introdução dos códigos de barra nos produtos comerciais
- Digital media
- CAD
- Data streams

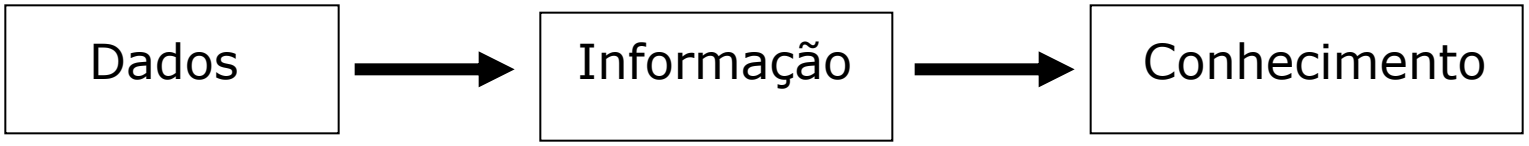
Formas de Análise dos Dados

- SQL - Linguagem de Interrogação de Bases de Dados
  - Sem capacidade de programação
  - Com limitações
- Folhas de Cálculo
- Bases de Dados Analíticas (**OLAP**)
  - Espaço de Dados Multidimensional
  - Versáteis, flexíveis – permitem combinar múltiplas dimensões de informação
  - Análises quantitativas dos dados

➥**Transformam Dados em Informação**

Motivação

- Desenvolvimento das capacidades informáticas
  - Novas formas de recolha de dados
  - Evolução na tecnologia de armazenamento de dados
    - ➥ Crescimento Exponencial do volume de dados
  - Aplicações mais complexas
  - Novos métodos de Análise de Dados
    - ➥ Análise Lógica versus Análise Gráfica



Dados, Informação, Conhecimento

- Primeiros SGBD's a ênfase recaía sobre conteúdo dos atributos das tabelas das BD's

  ➥ **dados**

- Estes dados passaram a ser manipulados através de ferramentas de análise de dados, SQL, Folhas de Cálculo, Sistemas EIS

  ➥ **informação**

- Novos métodos de Análise de Dados – baseados em técnicas de Inteligência Artificial, Estatística

  ➥**conhecimento**

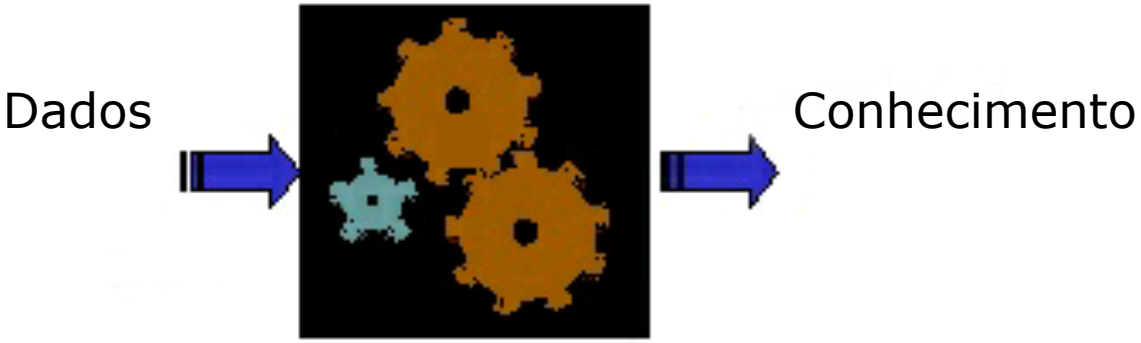
Necessidade de Conhecimento

Afogamo-nos em Informação

mas temos sede de conhecimento [Naibestt 1999]

É necessário extrair conhecimento interessante dos dados

Regras, Regularidades, Relações, Padrões



Data Science vs. Estatística

Estatística

Os processos são de **análise confirmatória** - um método só é aceite após a sua prova – teste da hipótese

- a distribuição dos dados tem de ser conhecida à partida
- o desenvolvimento e teste de uma hipótese é feito através do processo de análise
- as amostras são dados numéricos de reduzida dimensão

O termo Data Mining é conotado com análises *ad-hoc* que conduzem a descobertas de relações um pouco por acaso

Data Science

- Os processos são de natureza exploratória
- Para além de técnicas Estatísticas inclui técnicas de outras áreas: Inteligência Artificial, Bases de Dados, Ciências da Computação,...
- A maioria dos dados que ocorrem nas BD's são por natureza aleatórios, não lineares e de diferentes formatos e tipos (numérico, nominal, imagem, ...)
- Integra teoria e heurísticas, **assume-se uma atitude mais experimental**
- Foca-se nas várias fases do processo de descoberta de conhecimento: seleção, limpeza, integração e visualização dos resultados

Várias Designações

- Data Fishing, Data Dredging: 1960-
  usado pelos Estatísticos (como mau nome)

- Data Mining : 1990--
  usado pela comunidade das Bases de Dados, *Software houses*
  "database mining"™ - marca registada pela HNC

- Knowledge Discovery in Databases: 1995-
  usado pelas comunidades IA, Machine Learning
  Data Archaeology, Information Harvesting, Information Discovery,
  Knowledge Extraction, ...

- Atualmente: Data Science

Processo centrado em diferentes utilizadores

- **Especialista de domínio** possui amplo conhecimento da área em estudo

- **Analista** especialista no processo de DCBD e responsável pela sua execução. Este especialista deve conhecer profundamente todo o processo de descoberta de conhecimento e as técnicas mais adequadas a cada uma das suas fases

- **Utilizador final** usa o conhecimento extraído a partir do processo DCBD em aplicações que o auxiliam na tomada de decisões. Não é necessário que este utilizador tenha um conhecimento profundo da área em questão

O sucesso do processo de DCBD depende, em parte, da interação entre estes três tipos de utilizadores

A participação do especialista do domínio e/ou do utilizador final tem grande importância na **definição dos objetivos iniciais** do processo de DCBD, bem como na **avaliação final do conhecimento extraído**

Operações de Descoberta de Conhecimento

Aprendizagem Supervisionada

A aprendizagem é feita usando dados que estão etiquetados com a resposta correta. O algoritmo de ML aprende com os dados etiquetados e gera um modelo. Posteriormente, é dado ao modelo dados não etiquetados e o mesmo consegue prever qual o rótulo correto para esses dados

Aprendizagem não Supervisionada

Os dados não estão etiquetados e, portanto, o algoritmo atua sobre os mesmos sem orientação. Os algoritmos neste caso determinam semelhanças e padrões entre os dados

Operações de Descoberta de Conhecimento

Aprendizagem Semi-supervisionada

Apenas uma parcela dos dados está etiquetada

Aprendizagem por Reforço

É uma abordagem de treino dentro do ML, onde o agente é incentivado por meio de um sistema de recompensa que atribui valores positivos ou negativos com base no desempenho das suas ações.

Com o tempo, o agente gravita instintivamente em direção a ações que produzem resultados positivos, evitando aquelas que levam a consequências negativas

O Processo de Descoberta de Conhecimento

Processo de Descoberta de Conhecimento

É o processo não trivial de identificação de relações válidas, novas, compreensíveis e potencialmente úteis nos dados [Fayyad et al., 1995]

O conhecimento descoberto é usado para:

- Fazer classificações sobre novos dados
- Fazer previsões
- Sintetizar o conteúdo de grandes bases de dados
- Obter uma visão lógica dos dados

Conhecimento

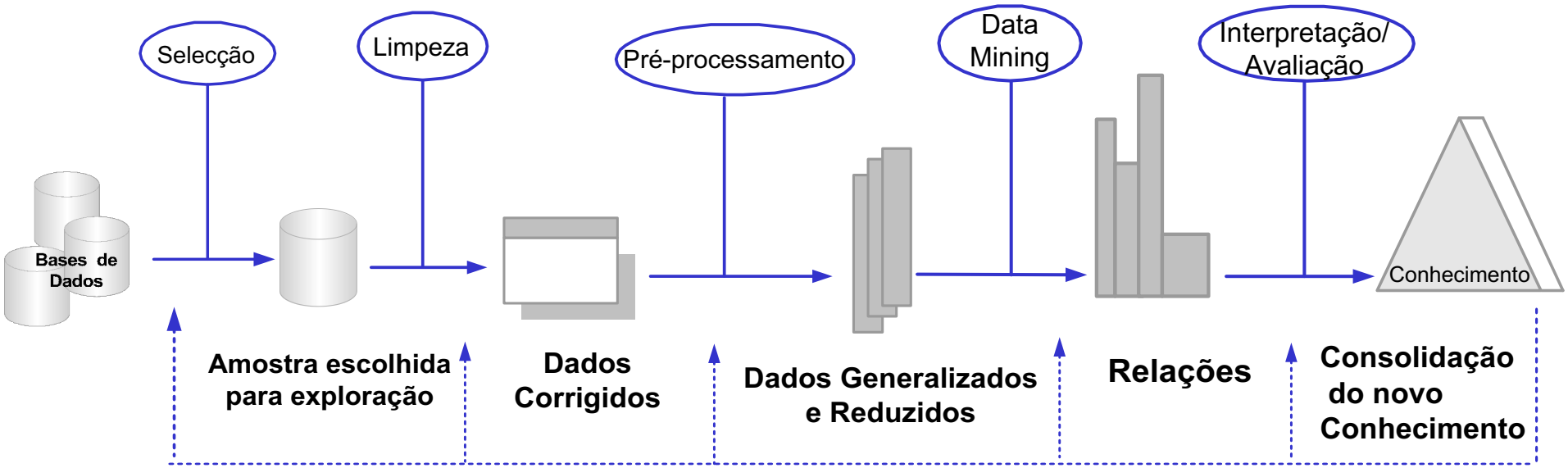
Sob a perspectiva de *Descoberta de Conhecimento,* o conhecimento é quantificado em termos de:

- Utilidade
- Validade
- Simplicidade/Complexidade
- Novidade

Estas medidas são aplicadas às relações/modelos

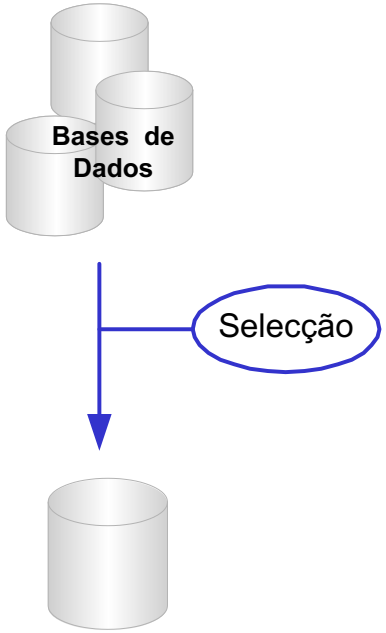
sempre sob a perspectiva de **Interesse**

Processo de Descoberta de Conhecimento



Processo Interactivo e Iterativo

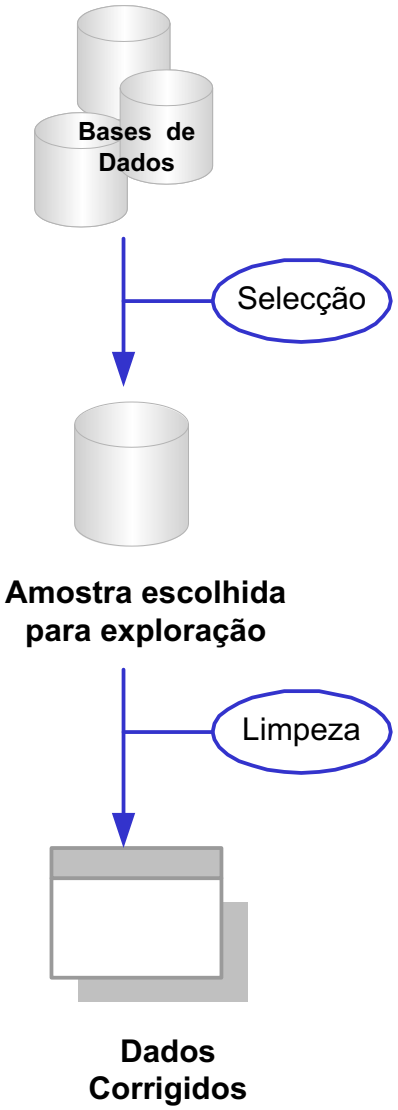
Fase de Selecção



Amostra escolhida para exploração

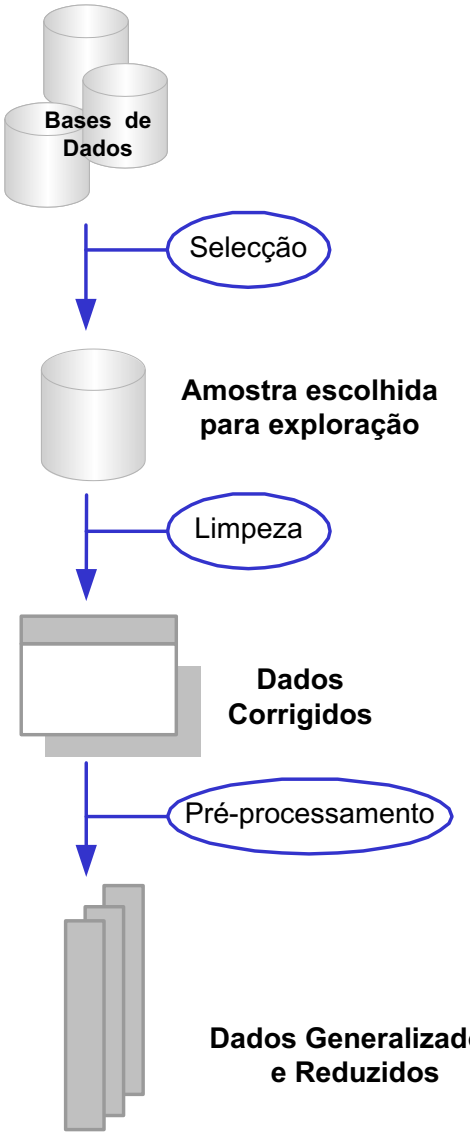
- Escolha dos dados de acordo com os objectivos de descoberta
- Volume de dados necessário
- Periodicidade de recolha das amostras
- Frequência de repetição dos exercícios de exploração

Fase de Limpeza



- Tratamento de dados em falta
- Tratamento de exemplos anormais
  - dados inconsistentes
  - valores isolados
- Eliminação de dados em mau estado
- Conversão de dados categóricos para valores numéricos
- Conversão de unidades

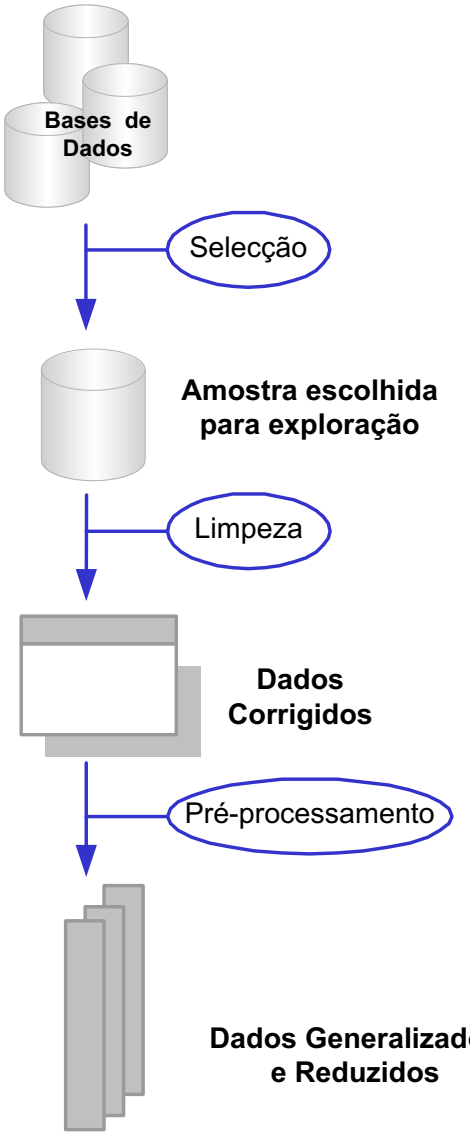
Fase de Pré-Processamento



Redução em Linhas

- Generalização de atributos categóricos
- Discretização de atributos contínuos
  - Algoritmos não sensíveis à classe
  - Algoritmos orientados por classes
- Normalização dos dados

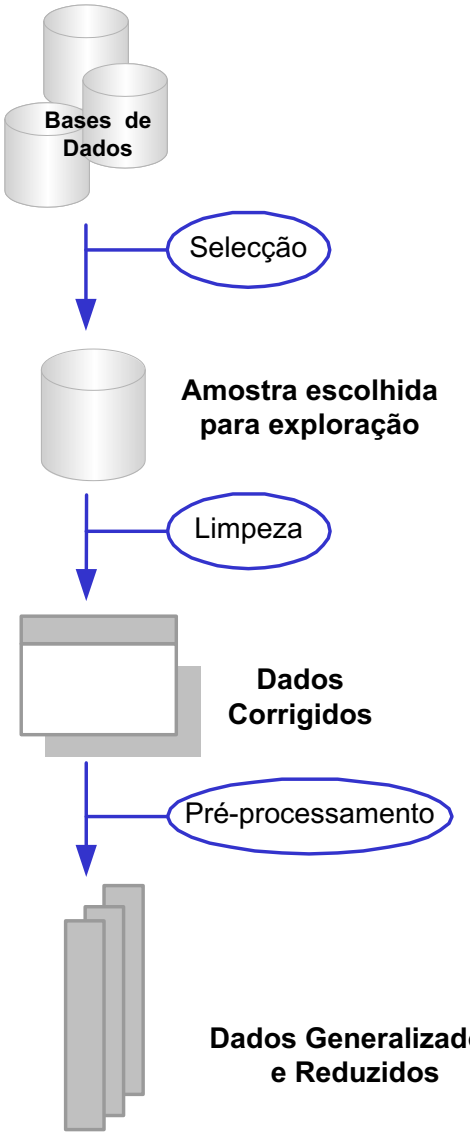
Fase de Pré-Processamento



Redução em Colunas

- Combinação de Variáveis de Entrada não correlacionadas
- Eliminação de variáveis correlacionadas
- Análise Sensitiva
- Análise dos Componentes Principais
- Aproximação Empacotadora
- Aproximação Filtro

Fase de Pré-Processamento

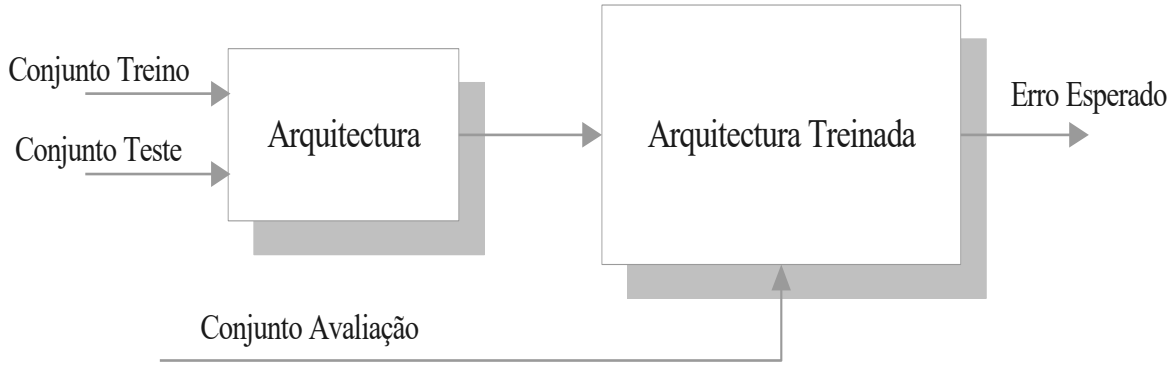


Sobre-ajustamento

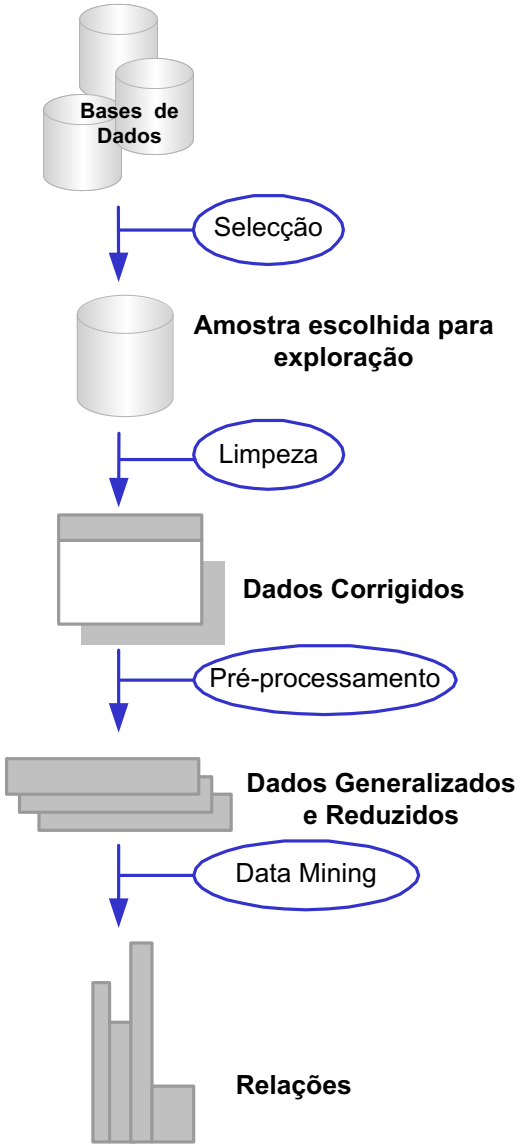
O modelo prevê os resultados baseado em particularidades dos dados usados no seu treino

Sub-ajustamento

O modelo falha na procura de relações de interesse nos dados, ou disponibiliza relações muito genéricas



Fase de Data Mining

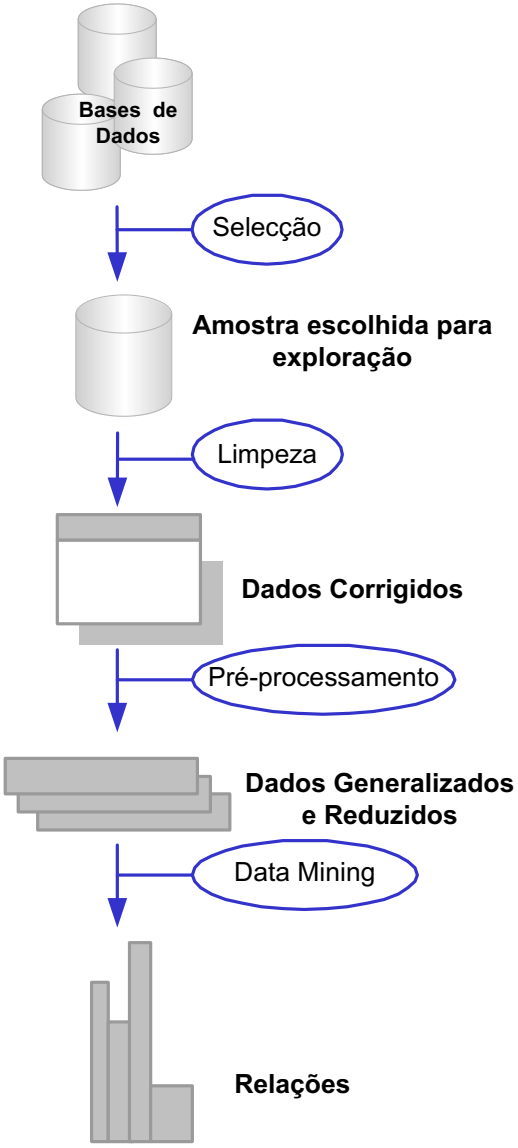


Envolve a adaptação de modelos, ou extracção de relações a partir dos dados, sem os passos adicionais que fazem parte de todo o processo de Descoberta de Conhecimento

Principais Operações de Data Mining

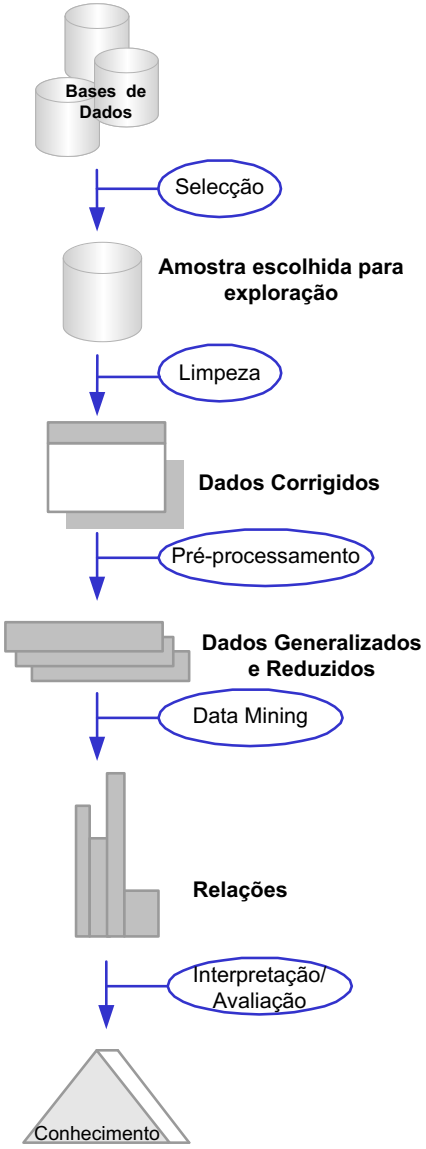
- Classificação
- Clustering
- Análise de Associações
- Análises Sequenciais
- Análise de Desvios
- Sumarização

Fase de Interpretação e Avaliação



- Visualização
- Filtragem de Conhecimento
  - Corte das regras
  - Limite mínimo de confiança das regras geradas
- Avaliação
  - Precisão
  - Taxa de Erro

Fase de Integração do novo Conhecimento

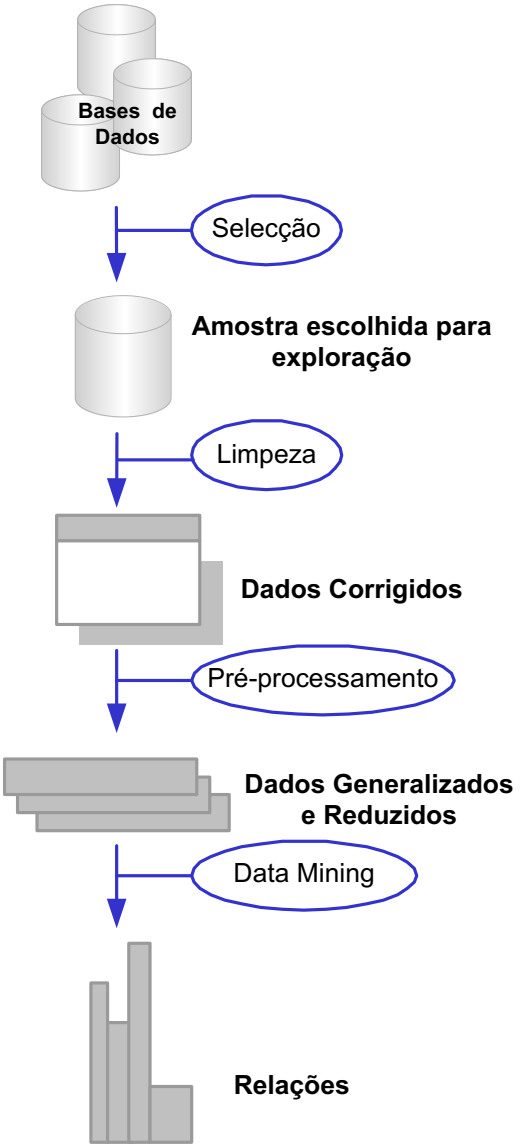


Integração do conhecimento num repositório central único pode envolver:

- modificação do conhecimento já existente (revisão)
- eliminação de conhecimento
- resolução de conflitos

Operações de Data Mining

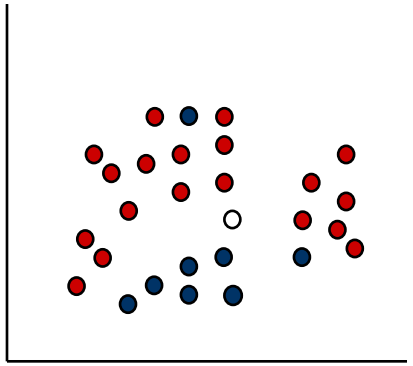
Classificação



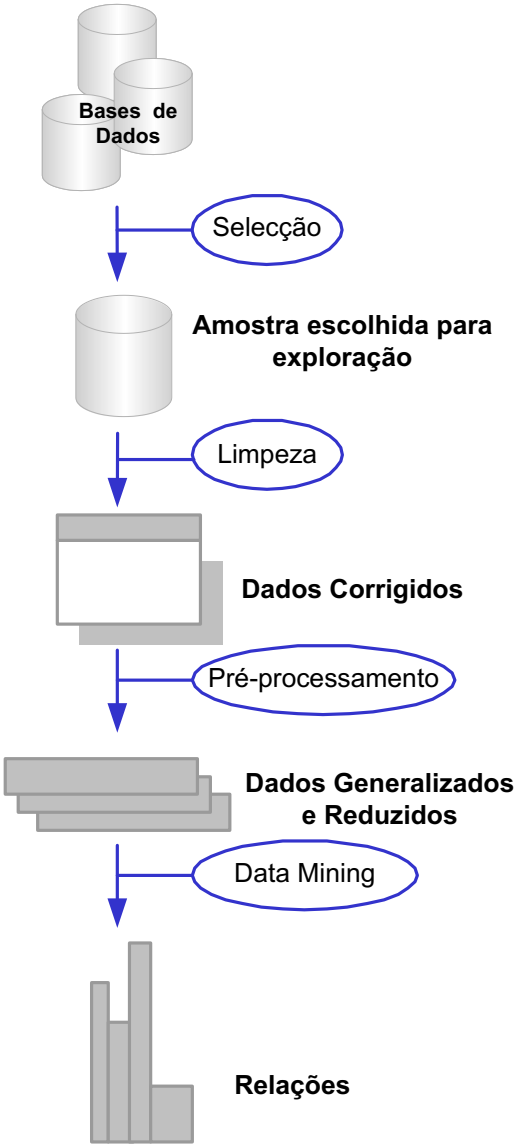
É uma função de aprendizagem que divide (ou classifica) os dados de acordo com um número específico de características.

Definição

- Seja uma base de dados $D=\{t_1,t_2,\ldots,t_n\}$
- um conjunto de classes $C=\{C_1,\ldots,C_m\}$,
- a Classificação consiste em definir uma relação $f:D\rightarrow C$ em que cada t_i é atribuído a uma classe
- a base de dados D é dividida em classes de equivalência



Classificação



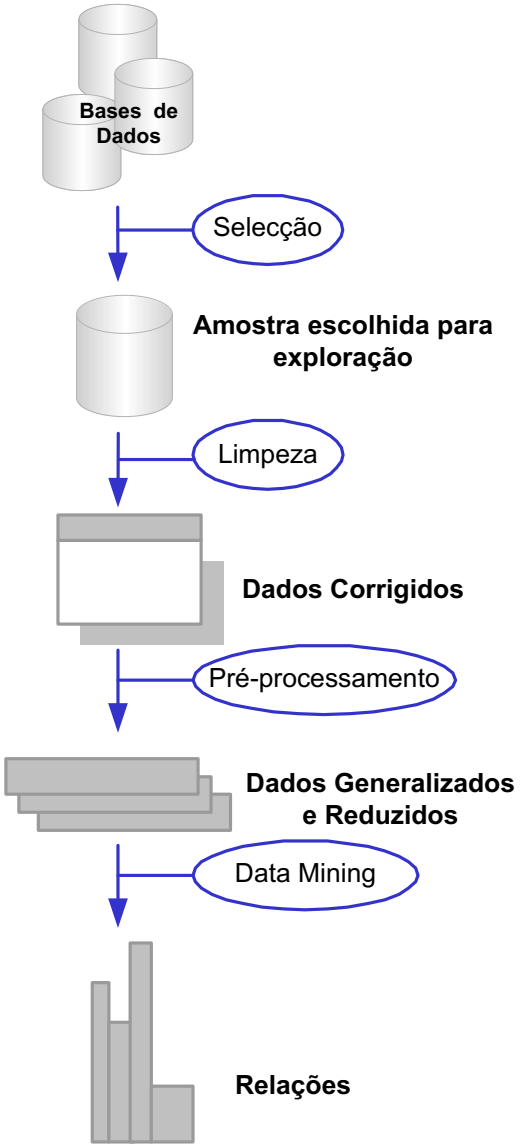
Aplicação

- Identificar potenciais clientes para uma campanha de marketing
- Identificar clientes com risco de crédito
- Reconhecimento de voz
- Reconhecimento de caracteres

Técnicas mais usadas:

- Árvores de Decisão
- Redes Neuronais
- Máquinas de Suporte Vectorial
- .....

Clustering

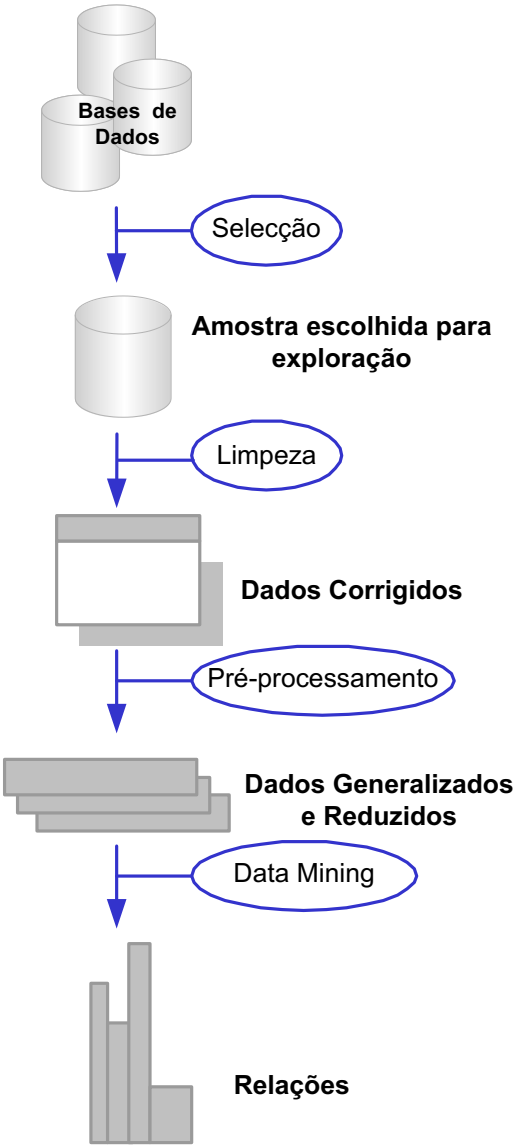


É uma operação que tem por objetivo identificar um conjunto finito de classes ou agrupamentos nos dados

Definição

- seja a base de dados D = $\{t_1,t_2,\ldots,t_n\}$
- um valor K (nº de classes)
- *Clustering* consiste em definir uma relação f:D→{1,..,k} em que cada tuple t_i é atribuído a um cluster K_j, 1<=j<=k
- o *Cluster*, K_j, contém precisamente os tuples a ele alocados
- o nº de clusters não é conhecido à priori

Clustering

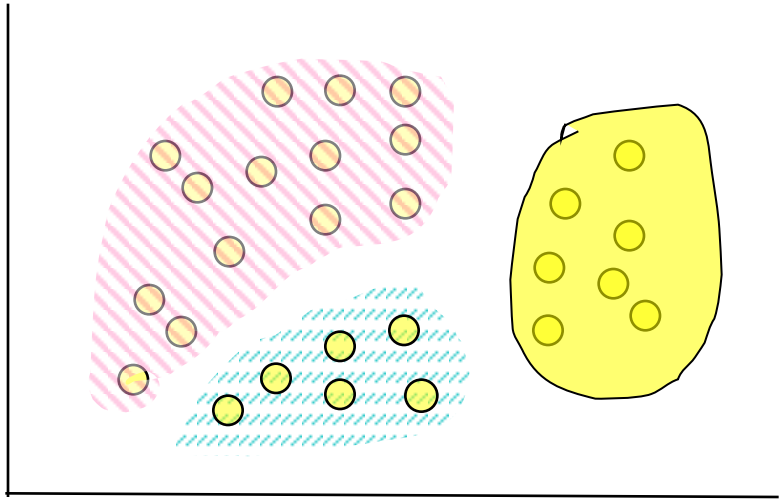


Aplicação:

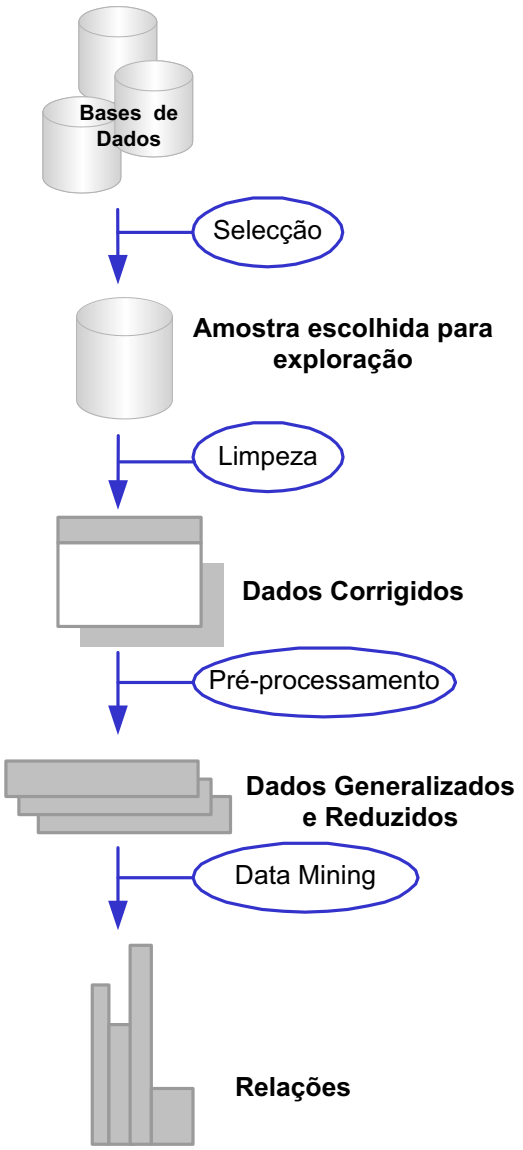
- sumariar o conteúdo de uma base de dados
- preparação de dados para outros métodos

Técnicas mais usadas:

- Técnicas Estatísticas - Algoritmo K-means
- Redes Neuronais
- Redes Kohonnen



Análise de Associações



Tem por objetivo gerar todas as associações entre items de transações que impliquem a presença de outros items

Definição

- seja um conjunto de itens $I=\{I_1,I_2,\ldots,I_m\}$
- uma base de dados de transações $D=\{t_1,t_2, \ldots, t_n\}$ em que $t_i=\{I_{i1},I_{i2}, \ldots, I_{ik}\}$ e $I_{ij} \in I$
- Análise de Associações consiste em identificar todas as regras de associação $X \Rightarrow Y$ com um Suporte e Confiança mínimo

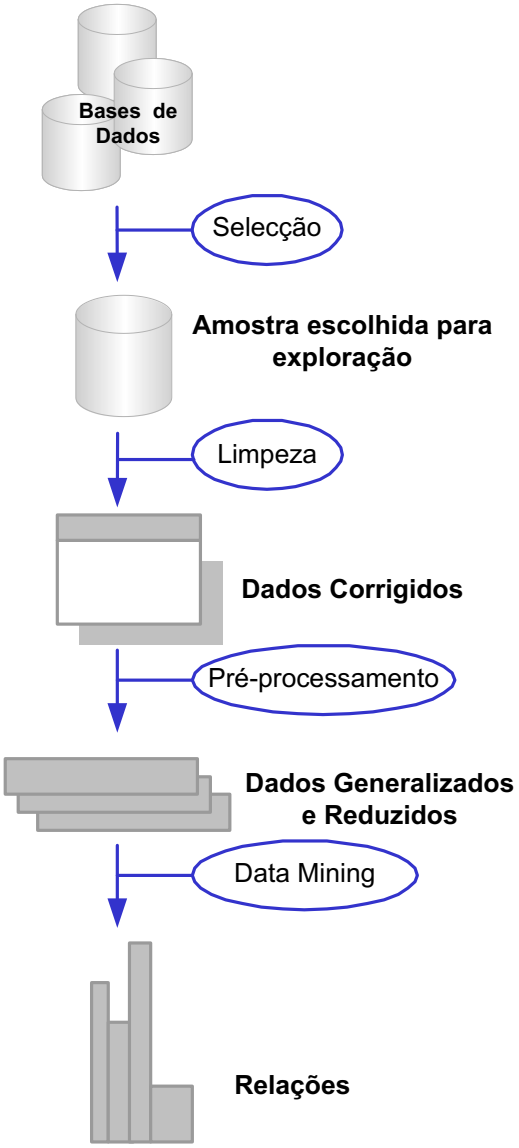
TID	Product
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	BREAD, BUTTER
9	MILK, BREAD, CEREAL, EGGS
10	MILK, BREAD, CEREAL

Conjuntos Itens Frequentes:

Milk, Bread (4)
Bread, Cereal (3)
Milk, Bread, Cereal (2)

Milk => Bread (Sup 40%, Conf 67%)

Análise de Associações



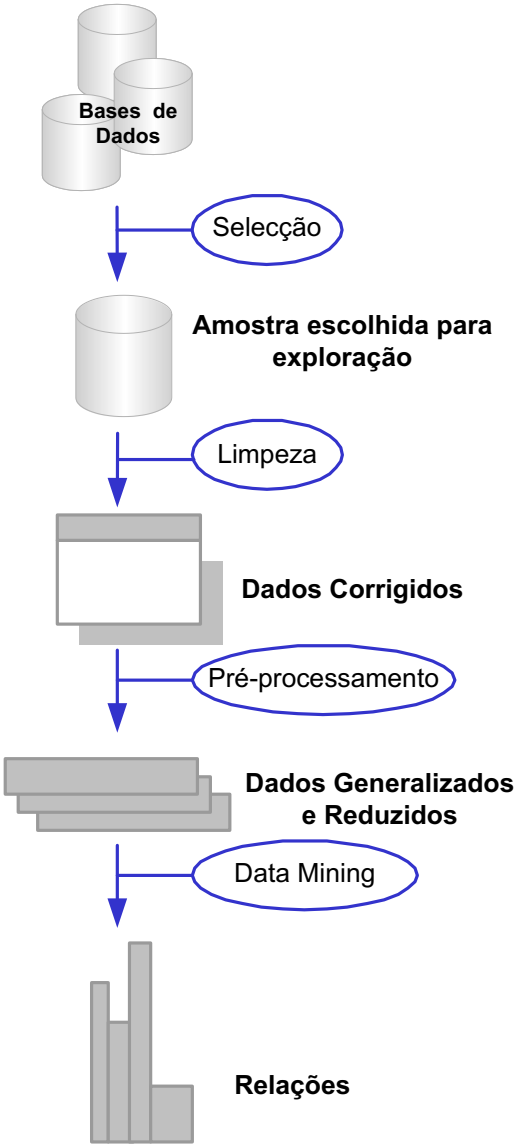
Exemplos

- Determinar produtos vendidos conjuntamente
- Relacionar diagnósticos médicos com valores de análises
- Relacionar acessos de páginas web

Técnicas mais usadas:

- Técnicas Estatísticas
- Algoritmo Apriori

Análises Sequenciais



Tem por objectivo encontrar regras que prevejam fortes dependências sequenciais entre diferentes eventos ao longo do tempo

- Regras são formadas primeiro por descoberta de padrões
- Ocorrências nas relações são ordenadas temporalmente

Exemplo

- Compras livrararia:

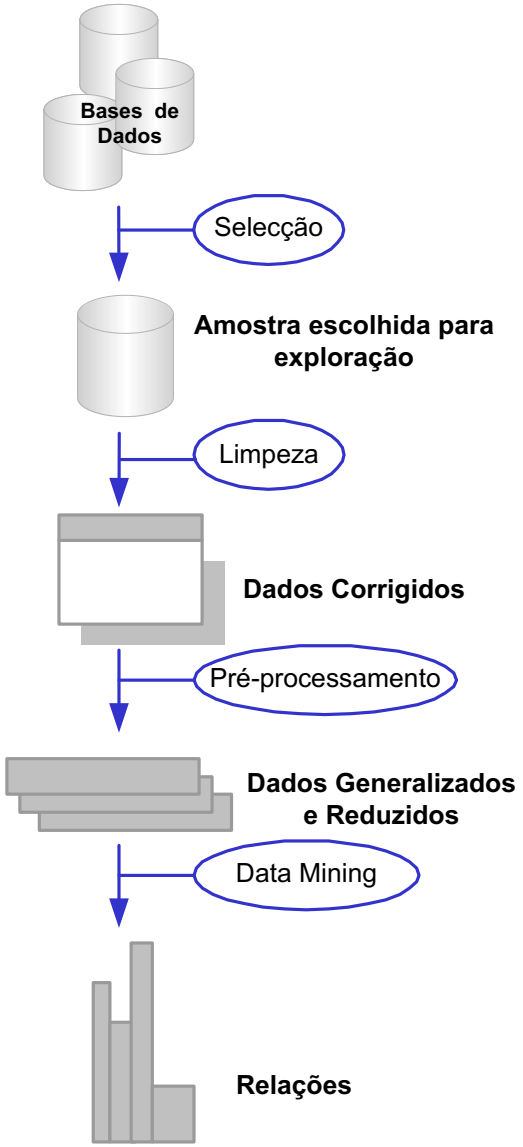
  (Intro_To_Visual_C) (C++_Primer) $\rightarrow$

  (Perl_for_dummies,Tcl_Tk)

- Alarmes de logs

  (Rectifier_Alarm) $\rightarrow$ (Fire_Alarm)

Análise de Desvios



Foca-se na descoberta de mudanças mais significativas nos dados a partir de valores previamente medidos ou valores normais

Exemplos

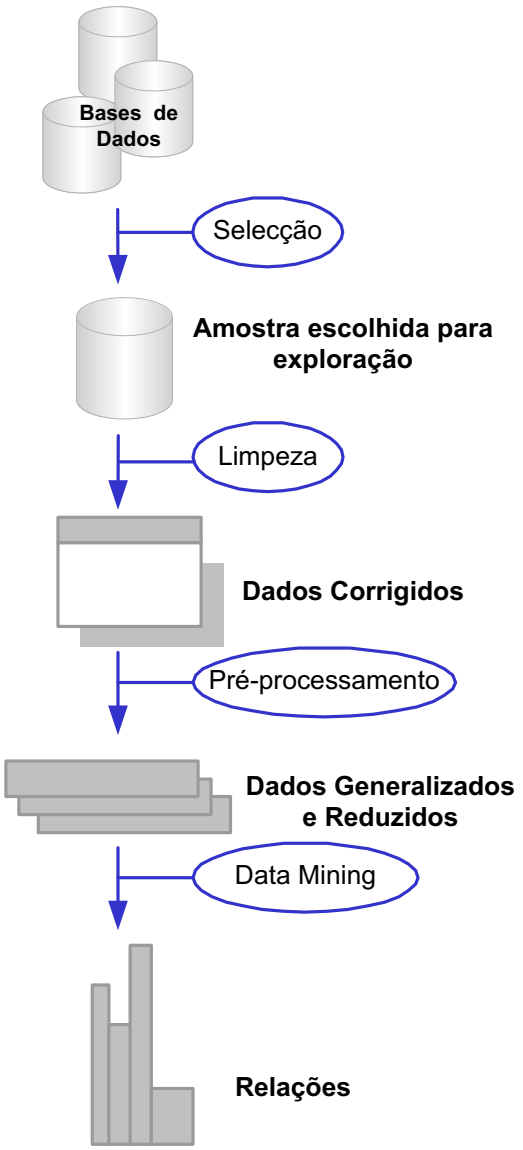
- Deteção de desvios em stocks
- Deteção de fraudes
- Deteção intrusos em redes

Exemplo: Agência de Viagens

- Os pacotes turísticos estão classificados em várias categorias: aventura, cultura, campo, praia...
- Cada pacote pode ter vários destinos, ou um só destino
- Sobre o mesmo pacote são feitos várias compras, por diferentes clientes em datas distintas
- A agência pode fazer promoções aos pacotes em vários períodos de tempo

Técnicas de Data Mining

Principais Técnicas de Data Mining

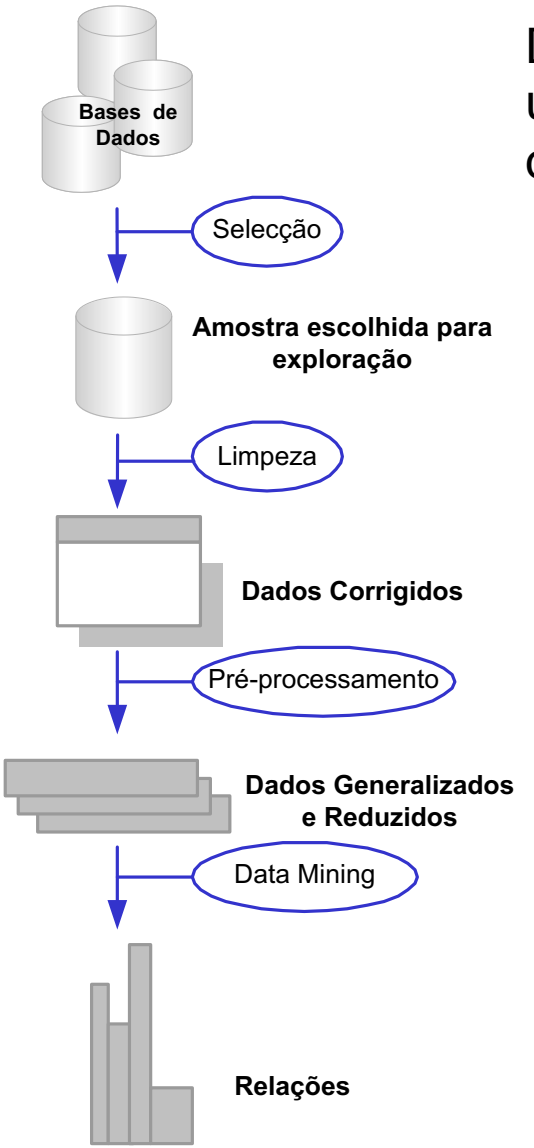


- Árvores de Decisão
- Redes Neuronais
- Máquinas de Suporte Vectorial
- Regressão
- Raciocínio Baseado em Casos
- Naive Bayes, Redes Bayesianas
- Algoritmos Genéticos
- .....

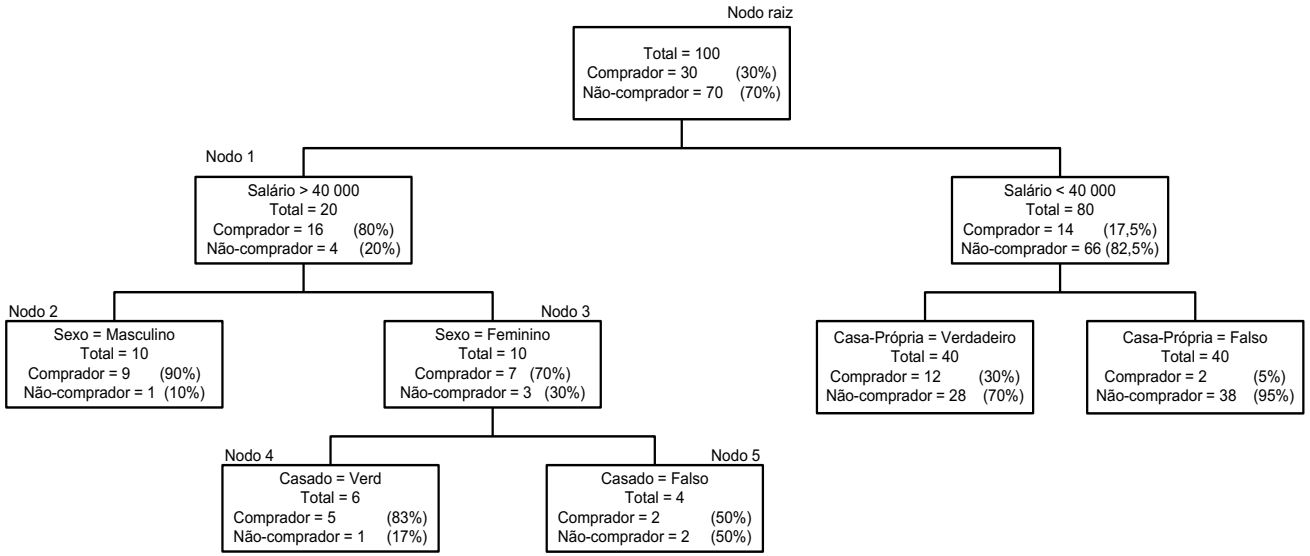
Uma técnica é adequada para fazer Data Mining:

- se produzir modelos de elevada qualidade
- se produzir modelos compreensíveis
- se puder aceitar conhecimento

Árvores de Decisão



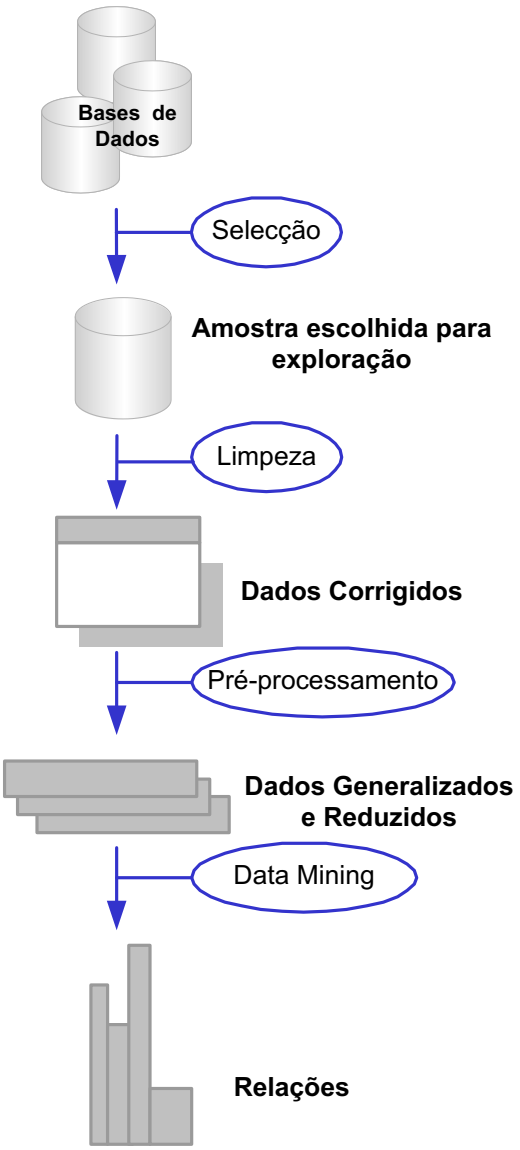
Dividem o conjunto de dados de modo a construir um modelo que classifica cada registo de acordo com o valor que apresentar no atributo objectivo



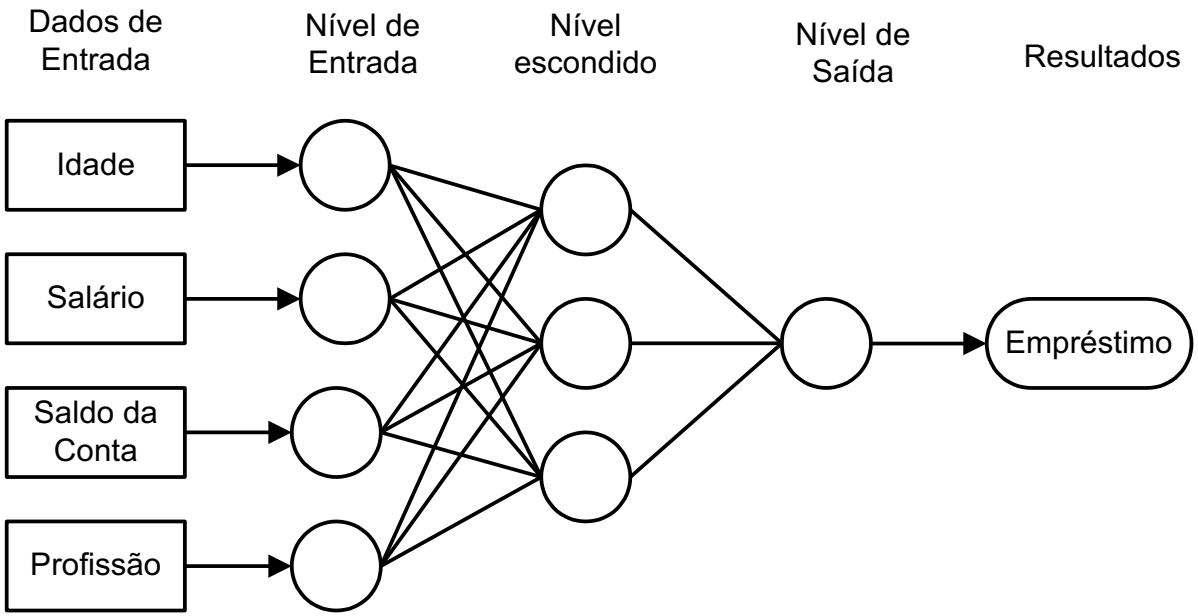
Algoritmos mais usados em ferramentas de Descoberta de Conhecimento:

- CART, C4.5

Redes Neuronais



São constituídas por uma série de nós interligados arranjados em níveis



Algoritmos mais usados em ferramentas de Descoberta de Conhecimento:

- Propagação Retroactiva – Classificação
- Função Base Radial – Classificação
- Rede Mapas Kohonen - Clustering

Top 10 Algoritmos Data Mining

- **C4.5** (indução de árvores de decisão e regras)
- **K-means** (Clustering)
- **SVM** (Máquinas de Suporte Vectorial)
- **Apriori** (Extração de Regras de Associação)
- EM (finite mixtures models)
- PageRank (Motor pesquisa Google, information retrieval)
- AdaBoost (combinação de classificadores)
- **kNN** (classificador baseado em instâncias)
- **Naive Bayes** (classificação baseado no Teorema de Bayes)
- **CART** (árvores de decisão)

Fonte original: "Top 10 algorithms in data mining", artigo da revista Knowledge and Information Systems, Dezembro de 2007

Teorema "*No Free Lunch*", Wolpert (1996)

Metodologias KDD

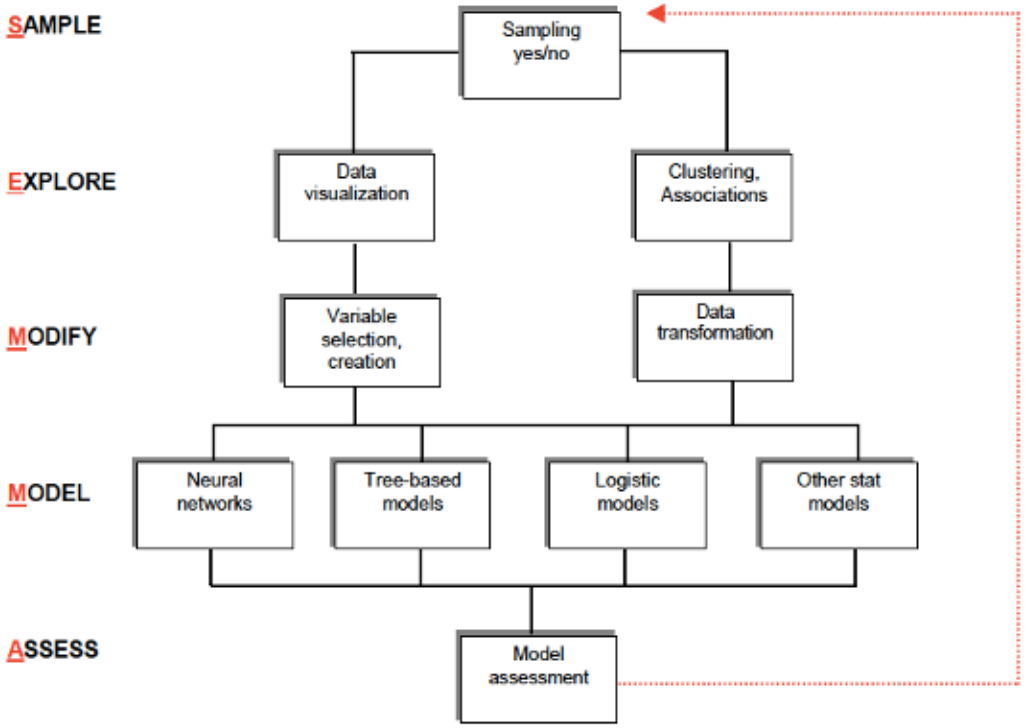
Com base no processo de DCBD foram definidas duas metodologias:

- **CRISP-DM** (CRoss Industry Standard Process for Data Mining)
  - Consórcio inicialmente composto com DaimlerChryrler, SPSS e NCR
- **SEMMA** (Sample, Explore, Modify, Model, Assessment)
  - SAS Enterprise Miner

As metodologias CRISP-DM e SEMMA são independentes das ferramentas, métodos ou técnicas de DM adoptadas, podendo ser usadas por qualquer uma

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

Metodologias KDD

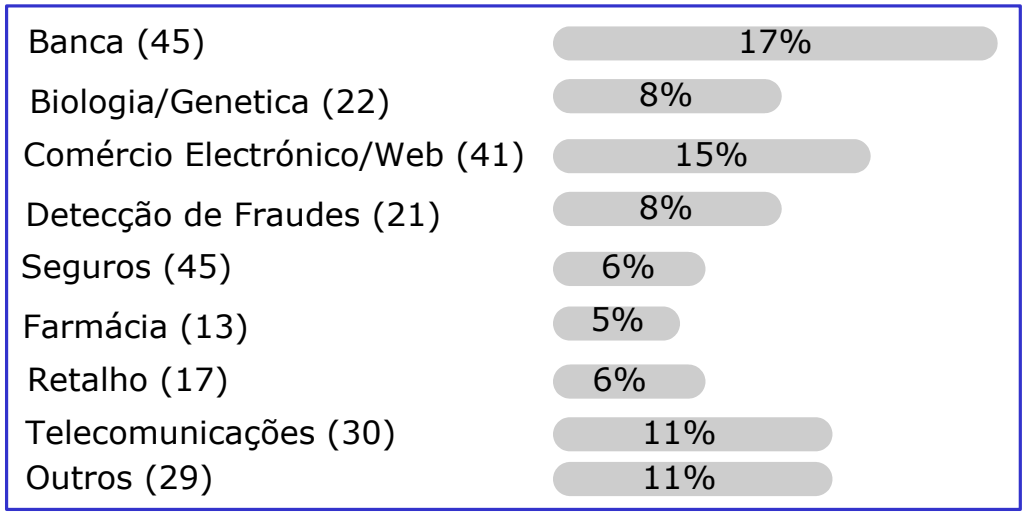


Metodologia SEMMA

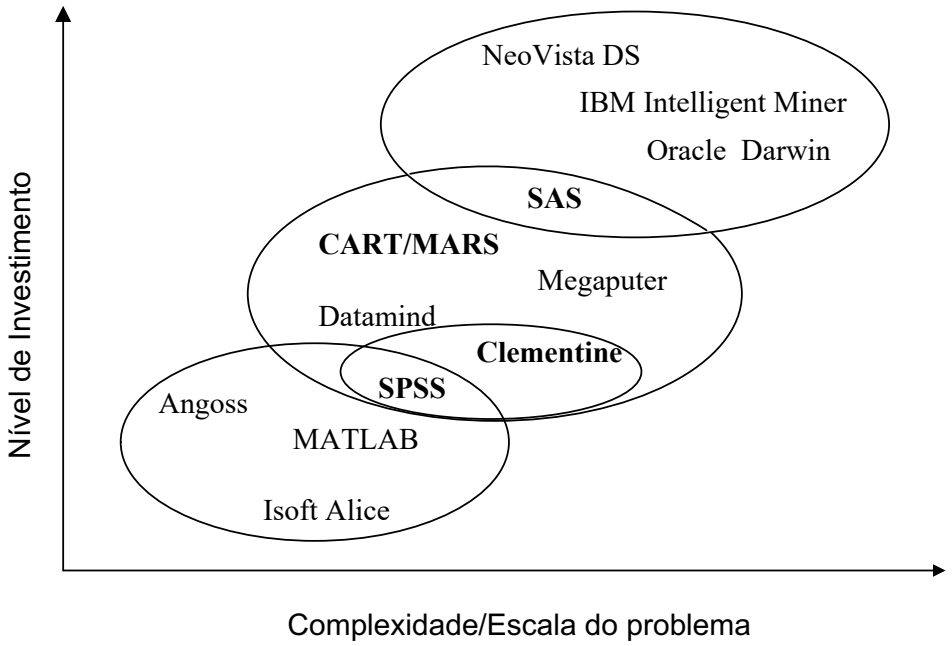
Metodologia CRISP-DM

Domínios de Aplicação

- Defesa
- Marketing&Vendas, Telecomunicações, Banca, Seguros
- Ciência & Medicina
- World Wide Web
- Text Mining
- Finanças
- Demografia
- Previsão de Audiências
- .....



Ferramentas por Segmentos Mercado



R (156)	18%
RapidMiner (135)	16%
KNIME (104)	12%
Weka / Pentaho (97)	11%
SAS (52)	6%
MATLAB (45)	5%
IBM SPSS Modeler (29)	3%
Orange (29)	3%
Outras (29)	11%

Inquérito realizado em Agosto 2017 no site www.kdnuggets.com

Top 3 open-source:

- Python
- R
- RapidMiner (Yale)
- …..

Big Data software you used in the past 12 months	
Apache Hadoop/Hbase/Pig/Hive (67)	8.4%
Amazon Web Services (AWS) (36)	4.5%
NoSQL databases (33)	4.1%
Other Big Data Data/Cloud analytics software (21)	2.6%
Other Hadoop-based tools (10)	1.3%

Data Mining versus Tipos de Dados

- **Text Mining:** Bases de dados textuais, e-mails, páginas web
- **Espacial Mining:** Sistemas de Informação Geográfica, Imagens
- **Multimedia Mining:** Bases de dados de imagem, vídeo/audio
- **Web Mining**
  - **Web Content Mining** - extrair conhecimento do conteúdo das páginas web (textos, gráficos, imagens, ...)
  - **Web Structure Mining** - extrair conhecimento da organização da Web, links entre referências, etc...
  - **Web Usage Mining** - também conhecida como Web Log Mining, extrair padrões interessantes dos logs dos servidores web
- **Reality Mining:** estuda interacções humanas com base no uso de dispositivos sem fio, como telefones celulares e sistemas de GPS

Data Mining e Ética

- Estará algum princípio ético ameaçado pela utilização exaustiva da análise de dados?

- Poderá o Data Mining contribuir negativamente para problemas de racismo, exclusão social, repressão ideológica?

    - Viés nos dados
    - Resultados Discriminatórios
    - Perfil social
    - Repressão através da Vigilância
    - Câmaras de eco e polarização
    - Exclusão da tomada de decisões
    - Violação da privacidade

Como mitigar estes problemas

- Garantir a qualidade dos dados
- Auditar regularmente modelos
- Diretrizes e Regulamentos Éticos
- Transparência e responsabilidade
- Equipes diversificadas e inclusivas
- Educar e aumentar a conscientização

Familiaridade/Interesse na área Data Mining

- Conhecem ou usam no dia a dia alguma aplicação de Data Mining?

- Têm algum problema a que potencialmente se possa aplicar Data Mining?

Alguns Apontadores

Sites

http://www.kdnuggets.com
(Maior site KDD: empresas, ferramentas, livros, publicações, conferências, empregos, ...)

www.acm.org/sigkdd
ACM SIGKDD – Associação profissional da sociedade Data Mining

Mailing Lists

http://www.kdnuggets.com/nuggets/index.html (KDD)
http://www.ics.uci.edu/~mlearn/MLList.html (Machine Learning)