



Introdução ao Processo de KDD e MD

Profa. Flavia Cristina Bernardini

KDD e MD

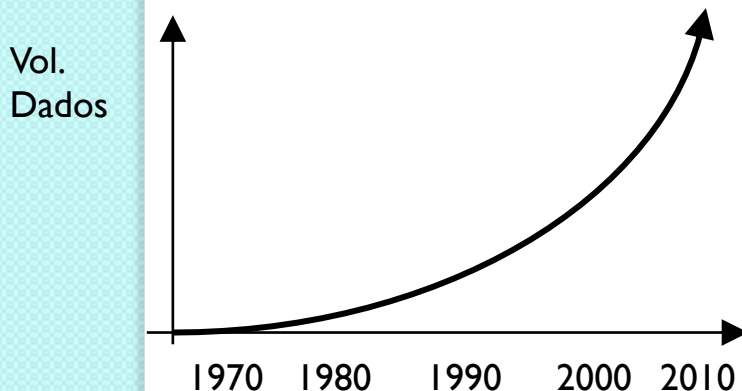
- KDD: Knowledge Discovery from Databases – Descoberta de Conhecimento de Bases de Dados
- MD: Mineração de Dados
- Conceitos propostos inicialmente por Fayyad, Piatetsky-Shapiro e Smyth em 1996
 - Leitura recomendada:
<http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>

Motivação

- “A quantidade de dados produzida e replicada no mundo dobra a cada 2 anos. [...] Até o fim de 2011, a previsão é que 1,8 zetabyte (ou 1,8 trilhão de gigabytes) de dados tenha sido criado e replicado”

Fonte:

http://chucksblog.emc.com/chucks_blog/2011/06/2011-idc-digital-universe-study-big-data-is-here-now-what.html



Por que precisamos de KDD?

- Método tradicional de análise dos dados – análise manual e interpretação
 - Saúde: especialistas analisam periodicamente tendências e mudanças no uso de serviços, através de relatórios detalhados
 - Geologia: geólogos analisam imagens remotas de planetas, asteroides, etc., localizando e catalogando objetos geológicos de interesse
- Análise clássica de dados:
 - Um ou mais analistas tornam-se familiares aos dados, servindo de interface entre os dados e os usuários dos dados (gestores e tomadores de decisão)

Definição

A Descoberta de Conhecimento em Bases de Dados (KDD) é o **processo interativo** para identificar nos dados **novos padrões** que sejam **válidos, novos, potencialmente úteis e interpretáveis**.

(Fayyad, Piatetsky-Shapiro e Smith, 1996)

Examinemos cada um destes termos:

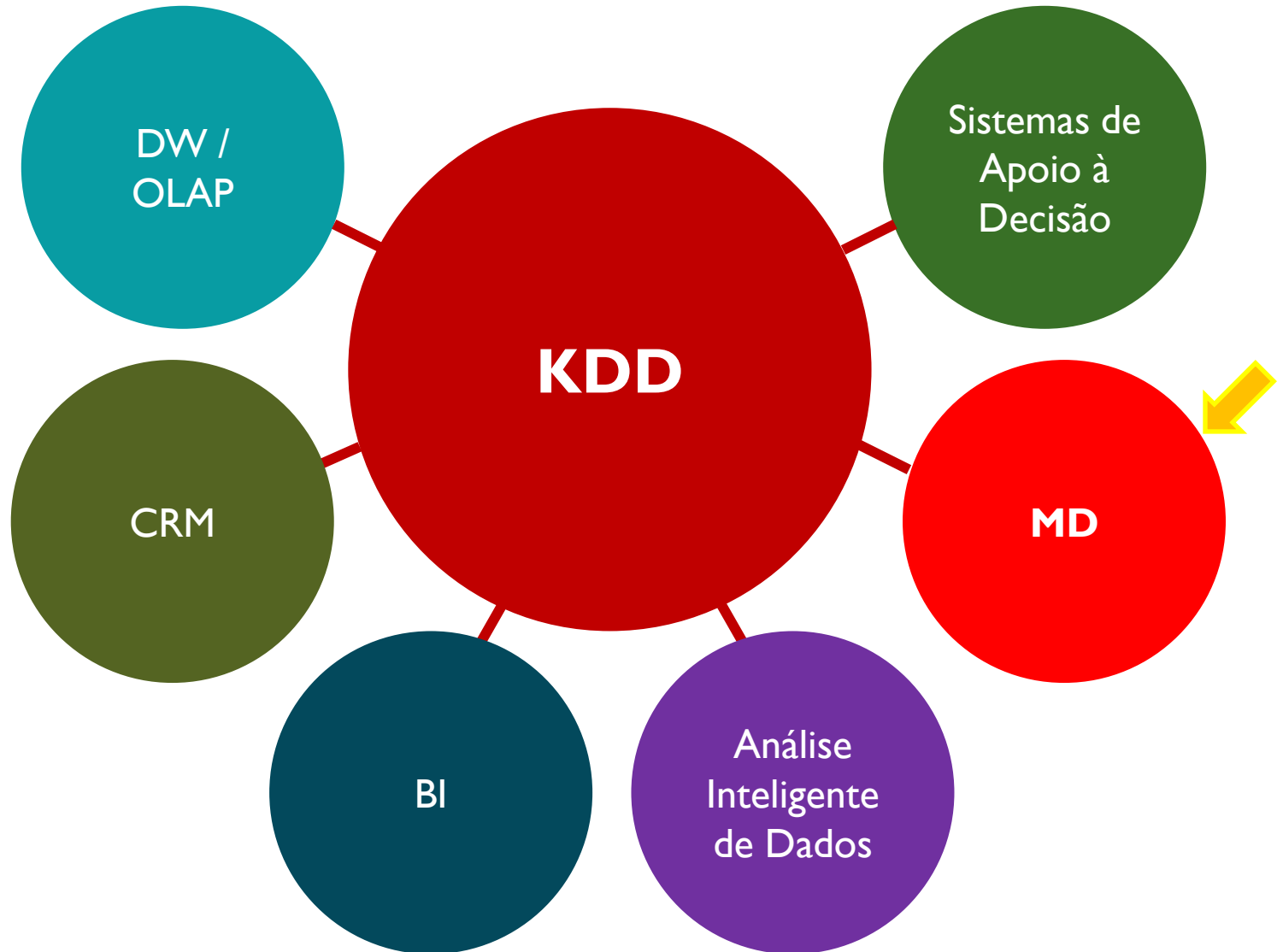
Definições

- Dados: Conjunto de fatos S
 - Ex: Tuplas de uma tabela atributo-valor
- Padrão: é uma expressão E numa linguagem L que descreve um sub-conjunto de fatos S_E do conjunto S :
 $S_E \subseteq S$
 - Ex: Dados sobre empréstimos bancários
 - O padrão E_1 = “Se Salário $< T$ então a pessoa falta ao pagamento” pode ser um padrão para uma escolha apropriada de T
 - Extrair um padrão também significa:
 - Ajustar um modelo aos dados
 - Encontrar alguma estrutura nos dados
 - Descrever em alto nível um conjunto de dados

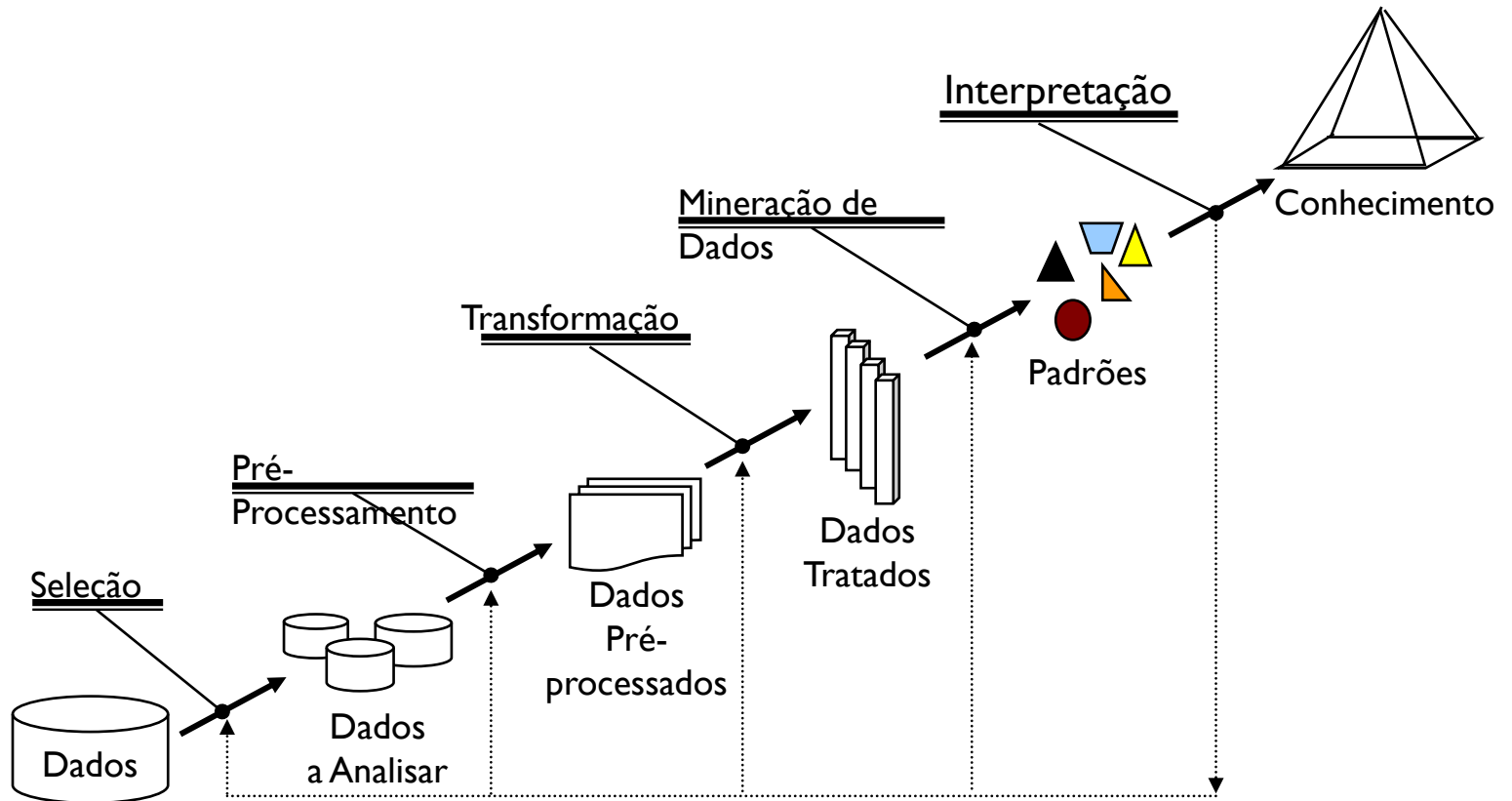
Exemplos

- Regras de atribuição de crédito para diminuir o crédito mal investido
- Perfil de Consumidor para um produto/campanha de marketing
- Associação entre grupos de clientes e compras
- Perfil de Fraude: cartões de crédito, subsídios...
- Previsão de tempo de internação de doentes em hospital
- Identificação de terapias/medicamentos de sucesso para uma doença
- Comportamento dos doentes para prever novas consultas
- Perfil de clientes que deixarão de usar uma empresa de telefonia celular
- Associação de usuários de um portal aos seus interesses/produtos
- Prever falhas numa máquina

Áreas Relacionadas



Processo



Antes de iniciar o processo...

- Compreensão do domínio de aplicação e do conhecimento prévio relevante
 - Uso de manuais, entrevistas com especialistas, etc

Leitura Recomendada para Aquisição de Conhecimento: GARCIA, A.C.B.; VAREJÃO, F.M.; FERRAZ, I.N. Aquisição de Conhecimento, Cap. 3. In: REZENDE, S.O. *Sistemas Inteligentes*. Ed. Manole, 2003.

- Identificação do objetivo do processo de KDD do ponto de vista do cliente

Fase: Seleção

- Aprendizado do domínio de aplicação.
- Seleção dos dados a analisar
 - agrupar os dados ou conjunto de variáveis sobre os quais se pretende trabalhar

Fase: Pré-processamento

- Pré-processamento e limpeza de dados:
 - operações básicas de remoção de ruído nos dados
 - decisão de estratégias em caso de campos omissos nos dados
 - consideração de sequências temporais nos dados

Fase: Transformação

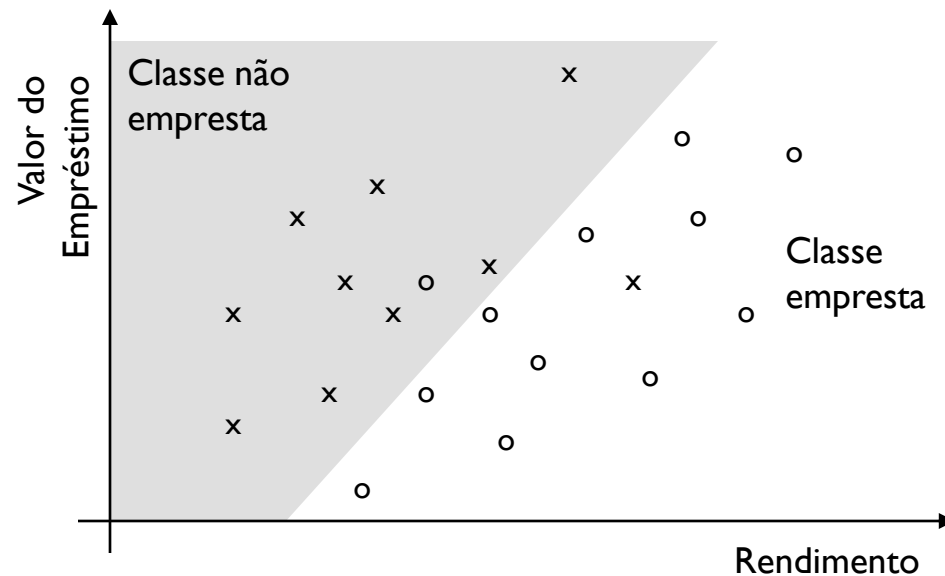
- Redução e Transformação dos dados:
 - procura de atributos úteis nos dados tendo em consideração os objetivos a que se destinam
 - utilização de métodos de transformação com vista à redução do número efetivo de variáveis em consideração
 - procura de representações invariantes para os dados

Fase: Mineração de dados (I)

- Adaptação dos dados para a tarefa de MD que se deseja
- Objetivos de MD:
 - Classificação: aprendizagem de uma função que mapeie os dados em uma ou várias classes
 - Regressão: aprendizagem de uma função que mapeie os dados em uma variável de previsão $y \in \mathbf{R}$
 - Clusterização/Segmentação: identificação de um conjunto finito de categorias ou clusters para descrição dos dados
 - Sumarização: utilização de métodos para procura de uma descrição compacta para um subconjunto de dados
 - Modelagem de Dependências ou Associações: busca por um modelo que descreva dependências significativas entre variáveis
 - Detecção de Alterações e Divergências: descoberta das alterações mais significativas nos dados a partir de valores medidos previamente ou normativos

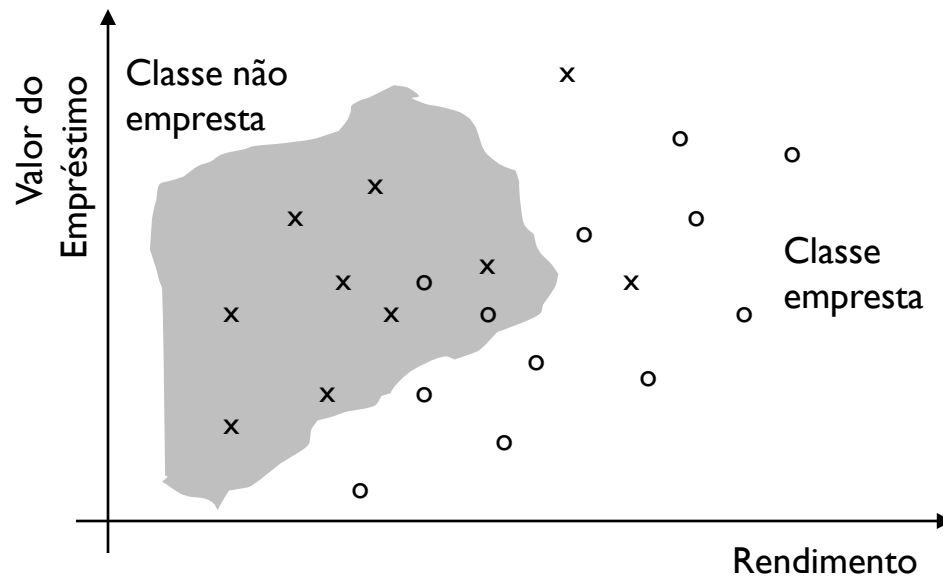
Fase: Mineração de dados (I)

- Exemplo de classificação linear
 - Necessidade de classificação prévia de ter havido (o) ou não (x) empréstimo para o cliente em questão



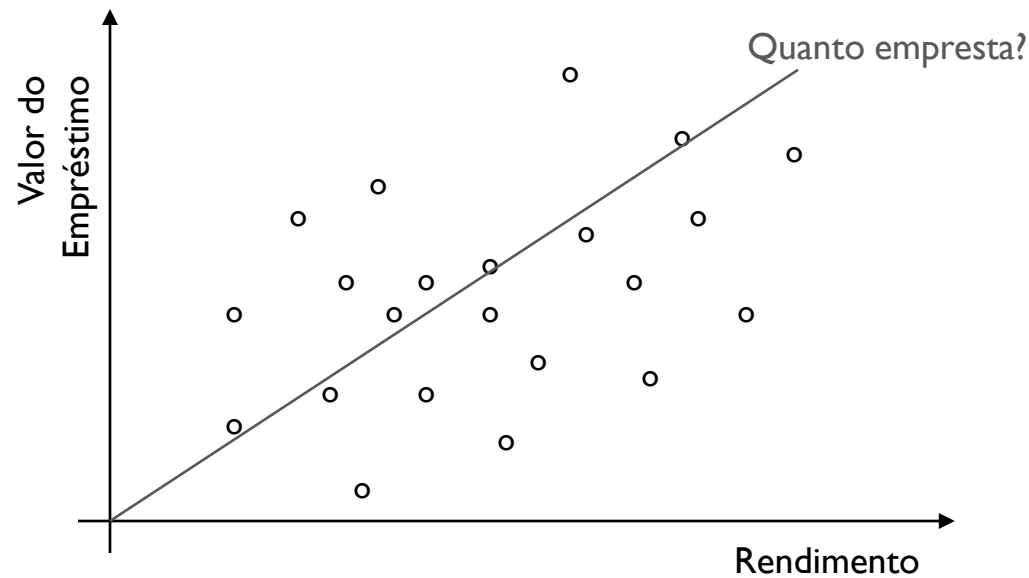
Fase: Mineração de dados (I)

- Exemplo de classificação não-linear
 - Ex: RNA, SVM, etc
 - Necessidade de classificação prévia de ter havido (o) ou não (x) empréstimo para o cliente em questão



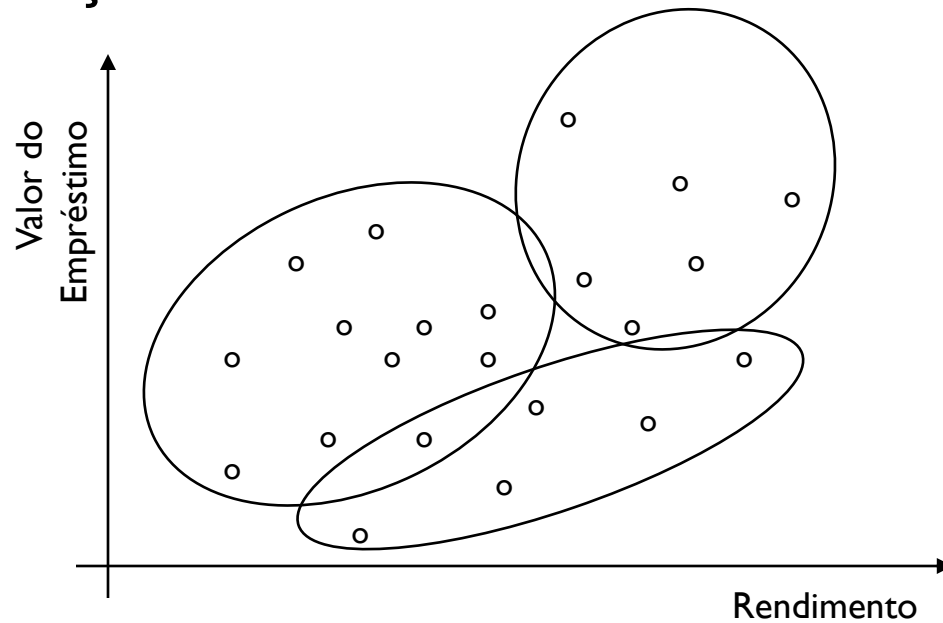
Fase: Mineração de dados (I)

- Exemplo de regressão linear
 - Necessidade de informação prévia dos valores que foram emprestados para cada cliente anteriormente



Fase: Mineração de dados (I)

- Exemplo de Clustering/Segmentação
 - Sem classificação prévia dos clientes
 - Exploração dos dados



Fase: Mineração de dados (II)

- Escolha do(s) algoritmo(s) de mineração de dados:
 - Árvores e Regras de Decisão (métodos indutivos)
 - Regressão Não Linear e Métodos de Classificação Não-Lineares
 - Algoritmos Genéticos Evolutivos
 - Redes Neurais Artificiais
 - Máquinas de Vetor Suporte
 - Métodos Baseados em Exemplos
 - K-Nearest Neighborhood – K-Vizinhos Mais Próximos
 - Modelos Gráficos Probabilísticos de Dependências
 - Redes Bayesianas
 - Naive Bayes
 - Modelos de Aprendizagem Relacional
 - Programação Lógica Indutiva

Fase: Interpretação

- Interpretação dos padrões minerados com o possível regresso a uma das fases anteriores para maior interação ou documentação
- Consolidação do conhecimento descoberto:
 - incorporação deste conhecimento no sistema ou elaboração de relatórios para as partes interessadas
 - Verificação e resolução de potenciais conflitos com conhecimento tido com verdadeiro (ou previamente extraído)

Problemas na Mineração de Dados

- Informação limitada
 - Informação incompleta
 - Dados dispersos
 - Espaço de testes
- Dados corrompidos
 - Ruído
 - Dados omissos
- Bases de Dados
 - Tamanho
 - Voláteis

Bibliografia

- FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery. *AI Magazine*, 1996, p. 37-54.
- FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. E UTHURUSAMY R. (Eds.). *Advances in Knowledge Discovery and Data Mining*. The MIT Press, Massachusetts, 1996.

Referências Bibliográficas

- FACELLI, K.; LORENA, A.C.; GAMA, J.; CARVALHO, A.C.P.L.F. *Inteligência Artificial – Uma Abordagem de Aprendizado de Máquina*. Ed. LTC, 2011.
- REZENDE, S.O. *Sistemas Inteligentes: Fundamentos e Aplicações*. Ed. Manole, 2003.
- WITTEN, I.H.; FRANK, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann (2ª Ed.) , 2005.
- ELMASRI, R.; NAVATHE, S.B. *Sistemas de Banco de Dados*. Addison-Wesley (6ª Ed.), 2011. (Disponível na biblioteca a 4ª Ed.)
- MITCHELL, T. M. *Machine Learning*. McGraw-Hill, 1997
- REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. *Revista de Sistemas de Informação da FSMA*, n. 7, 2011, pgs. 7-21. Disponível em http://www.fsma.edu.br/si/edicao7/FSMA_SI_2011_1_Principal_3.pdf

Recursos na Internet

- KDNuggets – Recursos da comunidade de Mineração de Dados:
 - <http://www.kdnuggets.com/>
 - Bases de Dados: <http://www.kdnuggets.com/datasets/index.html>
 - Competições: <http://www.kdnuggets.com/competitions/>
- Repositório de bases de dados – UCI:
 - <http://archive.ics.uci.edu/ml/>
 - Para KDD: <http://kdd.ics.uci.edu/> – grandes bases de dados
- Competições em KDD:
 - <http://www.kdd.org/kddcup/index.php>
- Kaggle – From big data to big analysts
 - <http://www.kaggle.com/>
 - Competições: <http://www.kaggle.com/competitions>
- Portal Brasileiro de Dados Abertos
 - <http://dados.gov.br/>

Objetivo da Disciplina

- Oferecer uma visão geral do Processo de Mineração de Dados
- Trabalhos:
 - Escolher uma base de dados para abordar
 - Recomenda-se a utilização de uma das bases disponíveis previamente tratadas
 - Dar preferência por bases que conheçam o domínio



Mineração de Dados

Introdução

- **BI**
 - Habilidades, tecnologias, aplicações e práticas para adquirir melhor entendimento do contexto comercial dos negócios de uma empresa
 - Tornar dados em informações úteis para usuários de negócios
 - Gerenciamento da Informação
 - Suporte à Decisão
- **Data Warehouse (DW)**
 - OLAP direcionada a consultas de usuários
 - Padrões podem continuar escondidos

Introdução

- Necessidade de ferramentas de análise:
 - “Qual produto de alta lucratividade venderia mais com a promoção de um item de baixa lucratividade?”
- Técnicas de análise dirigidas por computador → Extração de Conhecimento

Data Warehouse



OLAP

| Idade | Motivo | Duração | Valor | Risco |
|-------|--------|---------|--------|-------|
| 45 | Carro | 36 | 10,000 | Baixo |
| 20 | Negoc. | 20 | 35,000 | Alto |
| 37 | Casa | 40 | 30,000 | Baixo |
| 29 | Carro | 24 | 25,000 | Alto |
| 66 | Mobil. | 10 | 7,000 | Alto |

OLAP

| Idade | Motivo | Duração | Valor | Risco |
|-------|--------|---------|--------|-------|
| 45 | Carro | 36 | 10,000 | Baixo |
| 20 | Negoc. | 20 | 35,000 | Alto |
| 37 | Casa | 40 | 30,000 | Baixo |
| 29 | Carro | 24 | 25,000 | Alto |
| 66 | Mobil. | 10 | 7,000 | Alto |

39,4

Média

OLAP

| Idade | Motivo | Duração | Valor | Risco |
|-------|--------|---------|--------|-------|
| 45 | Carro | 36 | 10,000 | Baixo |
| 20 | Negoc. | 20 | 35,000 | Alto |
| 37 | Casa | 40 | 30,000 | Baixo |
| 29 | Carro | 24 | 25,000 | Alto |
| 66 | Mobil. | 10 | 7,000 | Alto |

107,00

Somatório

OLAP

| Idade | Motivo | Duração | Valor | Risco |
|-------|--------|---------|--------|-------|
| 45 | Carro | 36 | 10,000 | Baixo |
| 20 | Negoc. | 20 | 35,000 | Alto |
| 37 | Casa | 40 | 30,000 | Baixo |
| 29 | Carro | 24 | 25,000 | Alto |
| 66 | Mobil. | 10 | 7,000 | Alto |

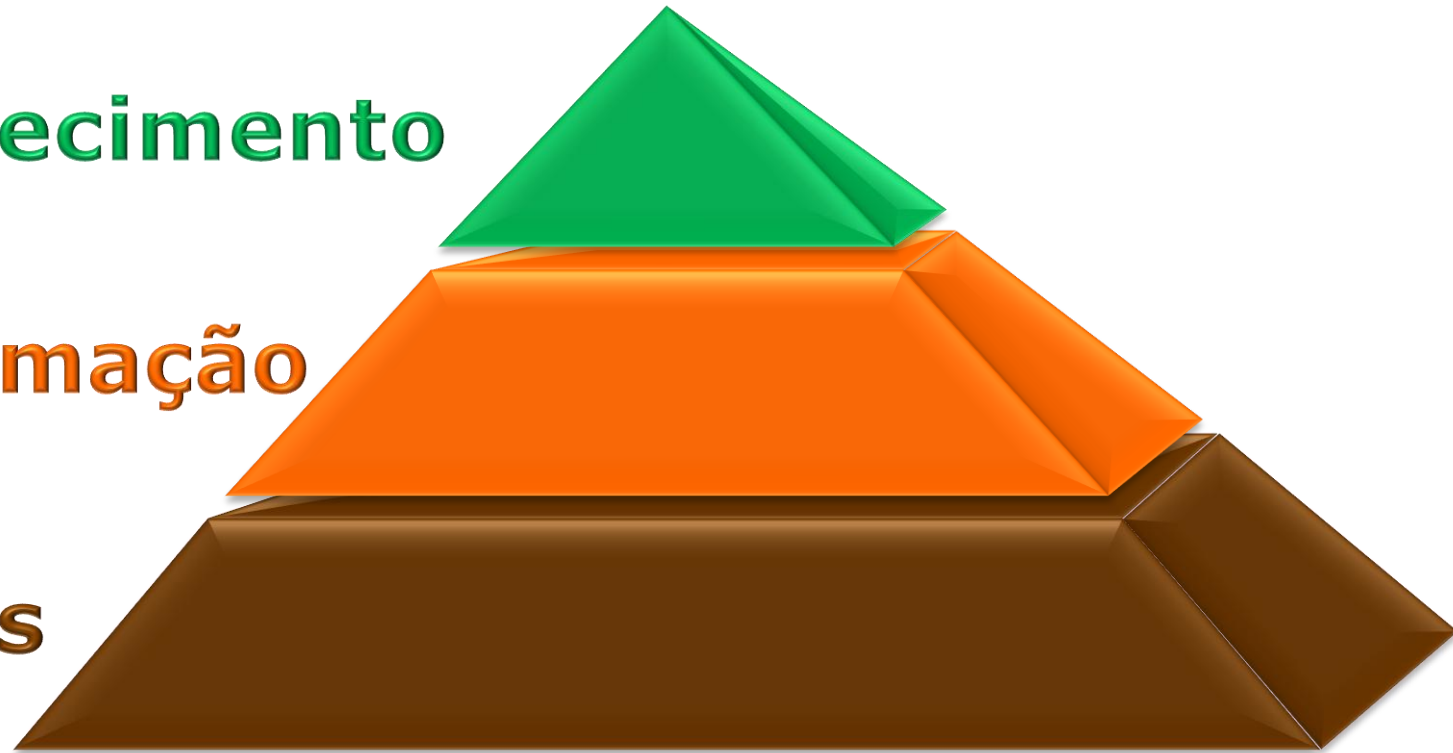
Seleção de uma parte dos dados

Estrutura

Conhecimento

Informação

Dados



Dado, Informação e Conhecimento

- **Dado:**

- É a estrutura fundamental sobre a qual um sistema de informação é construído
- Ex: Temperatura = 40°

- **Informação:**


- A transformação de dados em informação é freqüentemente realizada através da apresentação dos dados em uma forma compreensível ao usuário
- Ex: Febre alta = temperatura > 39°

Dado, Informação e Conhecimento

- **Conhecimento:**

- Fornece a capacidade de resolver problemas, inovar e aprender baseado em experiências prévias
- Uma combinação de instintos, idéias, regras e procedimentos que guiam as ações e decisões
- Se especialistas elaboram uma norma (ou regra), a interpretação do confronto entre o fato e a regra constitui um conhecimento
- Ex: Febre alta associada a cor amarela → hepatite

Mineração de Dados

- Objetivo: Extrair padrões de dados, em alguma linguagem de descrição
- Tipos de Mineração de Dados:
 - Mineração de Dados Visual
 - Métodos e Técnicas de IA que apoiam a MD 
 - ...

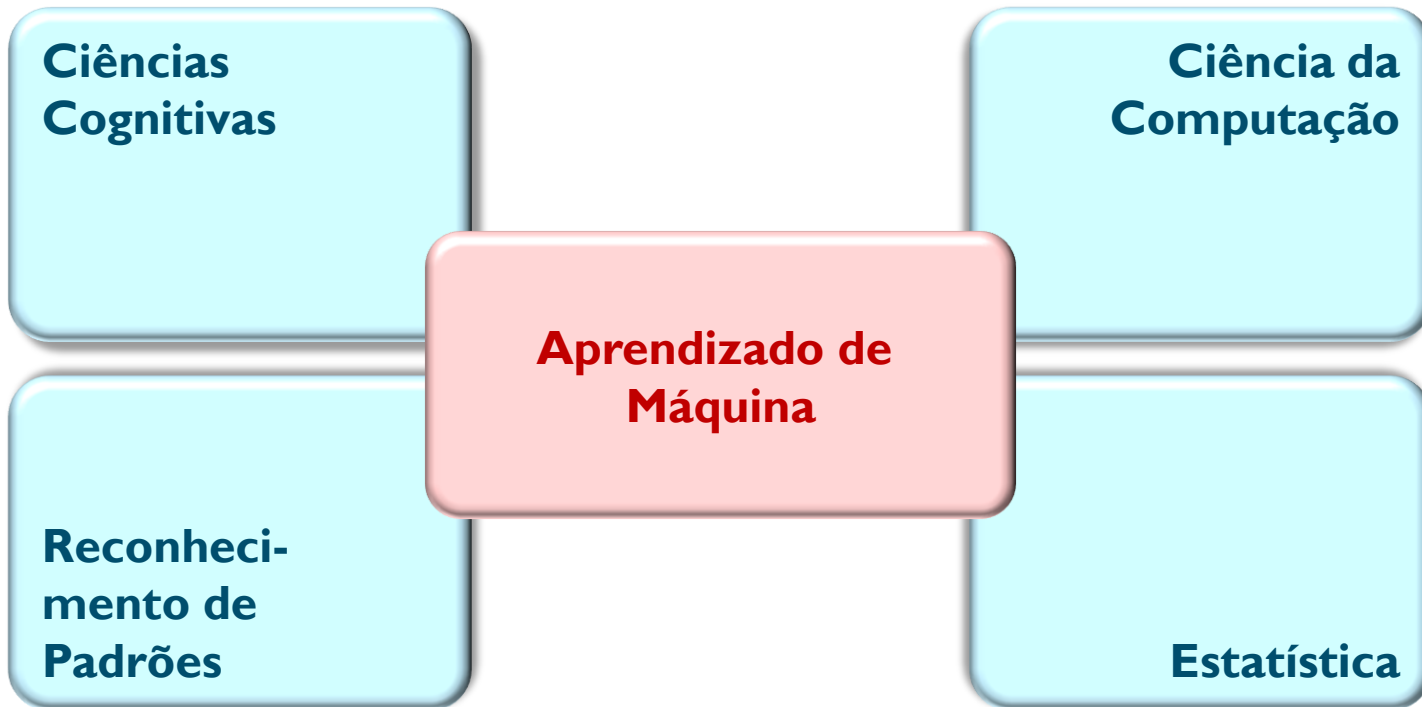
Aprendizado de Máquina

Sub-área da Inteligência Artificial que pesquisa métodos computacionais relacionados à aquisição de novos conhecimentos, novas habilidades e novas formas de organizar o conhecimento já existente

Objetivos de AM

- um melhor entendimento dos mecanismos de aprendizado humano
- automação da aquisição do conhecimento

AM incorpora várias técnicas de outras disciplinas



Tipos de Aprendizado

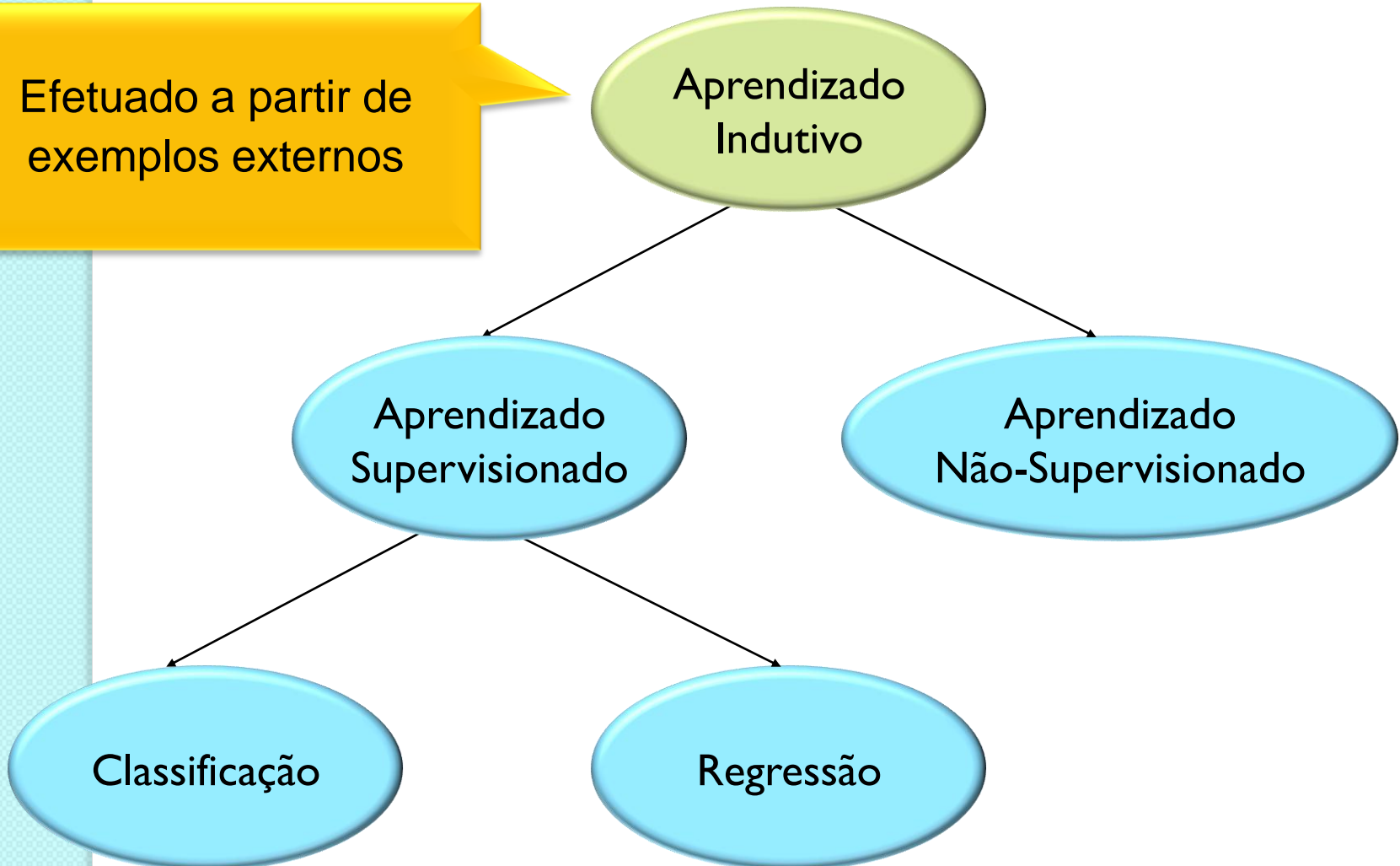
- **Aprendizado Indutivo:**
 - Indução é a forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos
 - O raciocínio é originado em de um conceito específico para um conhecimento genérico, ou seja, da parte para o todo
 - Na indução, um conceito é aprendido efetuando a inferência indutiva sobre os exemplos apresentados – construção de uma hipótese
 - Essa inferência pode ser verdadeira ou não
- **Obs:** Não existe um único algoritmo que apresente o melhor desempenho para todos os problemas – Teorema do No-Free Lunch

Tipos de Aprendizado

- Aprendizado por Reforço:
 - É Baseada em dados de um ambiente completamente observável
 - Inicialmente são conhecidos somente o estado inicial e final do problema
 - A cada etapa, é aprendido o próximo passo
 - É definido um sistema de recompensa para indicar se a decisão tomada para mudança de estado foi uma boa escolha

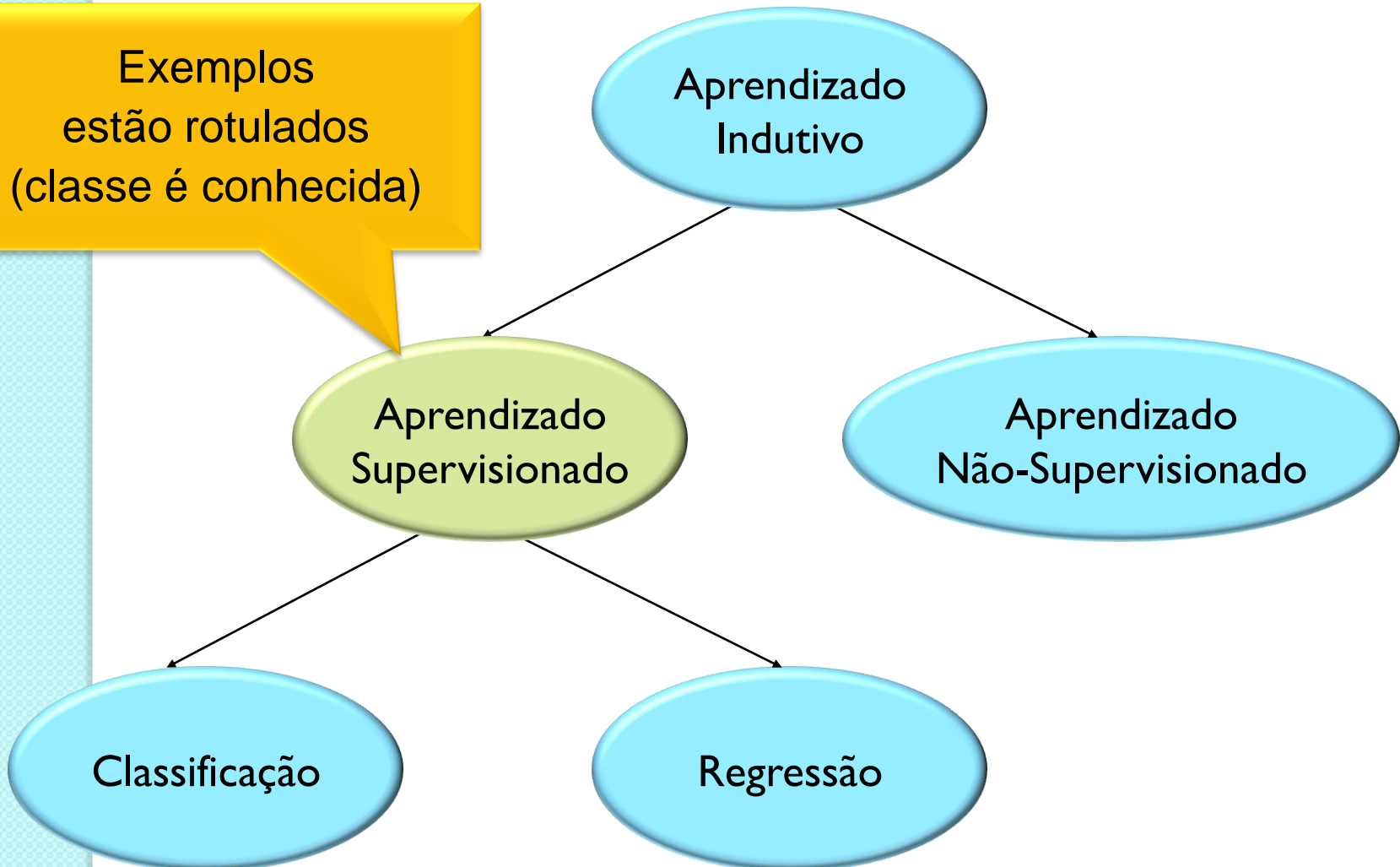
Hierarquia do Aprendizado

Efetuada a partir de exemplos externos



Hierarquia do Aprendizado

Exemplos
estão rotulados
(classe é conhecida)



Hierarquia do Aprendizado

Exemplos
sem rótulos
(classe desconhecida)

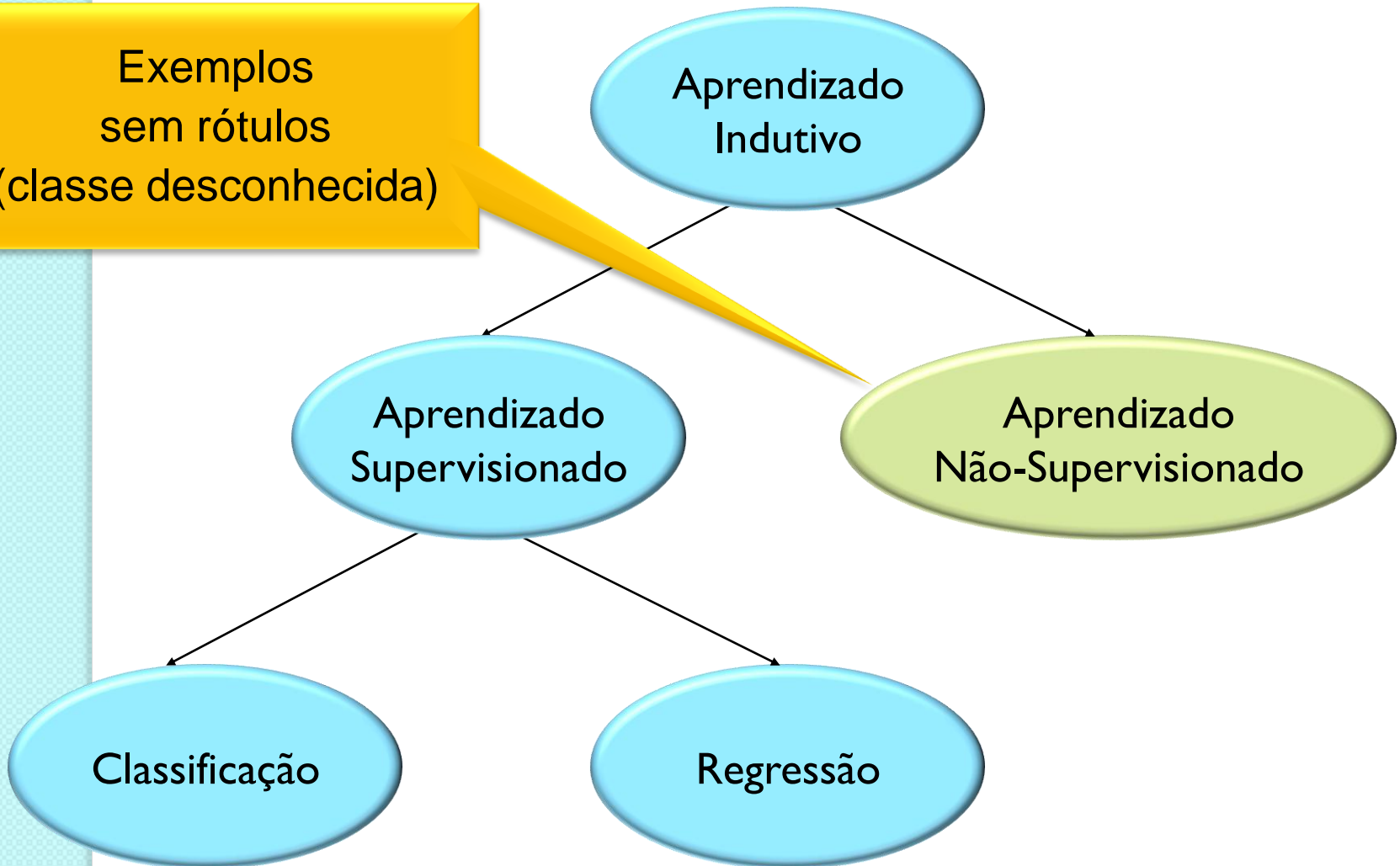
Aprendizado
Indutivo

Aprendizado
Supervisionado

Aprendizado
Não-Supervisionado

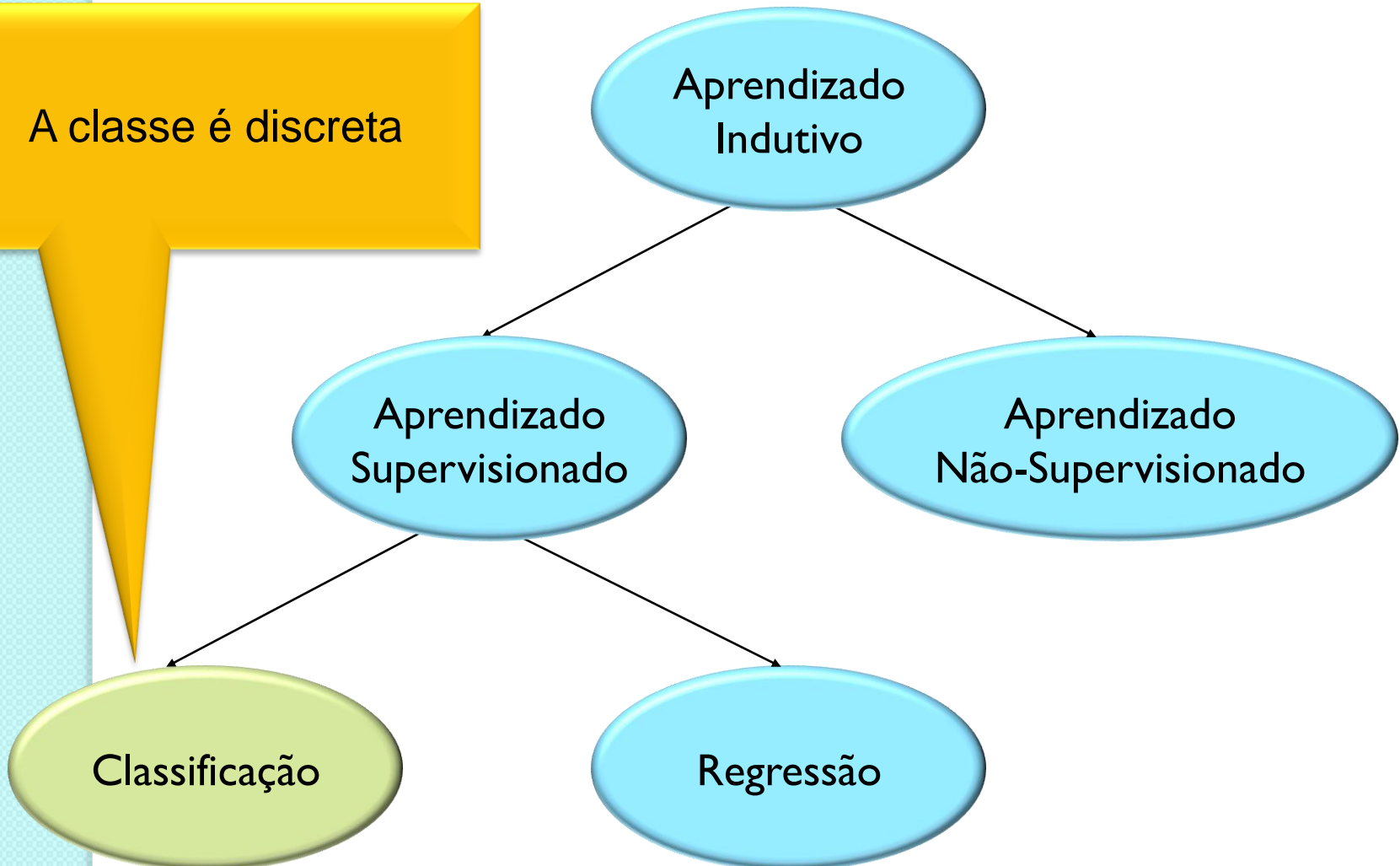
Classificação

Regressão



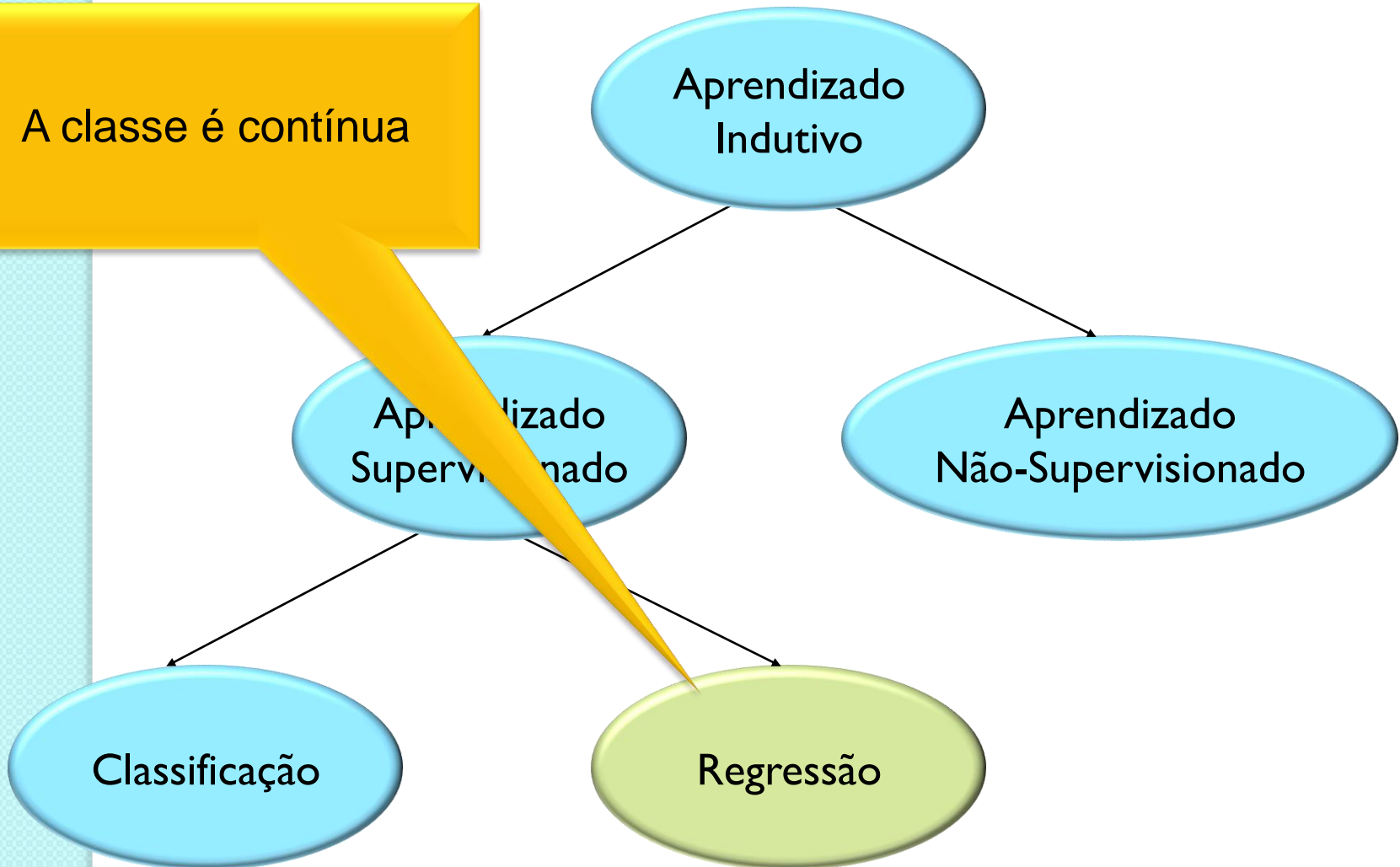
Hierarquia do Aprendizado

A classe é discreta



Hierarquia do Aprendizado

A classe é contínua



Aprendizado por Indução

Aprendizado Não Supervisionado:

Aprendizado por observação e descoberta



Aprendizado Supervisionado:

Aprendizado por exemplos

Sistemas de Aprendizado de Máquina

| Modo de Aprendizado | Paradigmas de Aprendizado | Linguagens de Descrição | Formas de Aprendizado |
|---|--|---|---|
| <ul style="list-style-type: none">○ Supervisionado○ Não Supervisionado | <ul style="list-style-type: none">○ Simbólico○ Estatístico○ Instance-Based○ Conexionista○ Genético | <ul style="list-style-type: none">○ Instâncias ou Exemplos○ Conceitos Aprendidos ou Hipóteses○ Teoria de Domínio ou Conhecimento de Fundo | <ul style="list-style-type: none">○ Incremental○ Não Incremental |

Processo de Classificação



Especificação do Problema

Conhecimento de Fundo



Conhecimento de Fundo

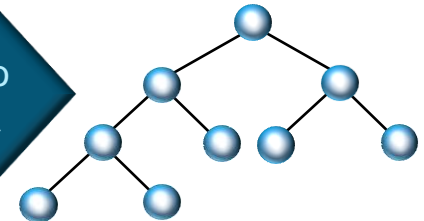
Variáveis Independentes

Variável Dependente

| | X ₁ | X ₂ | X ₃ | X ₄ | Y |
|----------|----------------|----------------|----------------|----------------|---------|
| sunny | 25 | 72 | yes | | go |
| sunny | 28 | 91 | yes | | dont_go |
| sunny | 22 | 70 | no | | go |
| sunny | 23 | 95 | no | | dont_go |
| sunny | 30 | 85 | no | | dont_go |
| overcast | 23 | 90 | yes | | go |
| overcast | 29 | 78 | no | | go |
| overcast | 19 | 65 | yes | | go |
| overcast | 26 | 75 | no | | go |
| overcast | 20 | 87 | yes | | dont_go |
| rain | 22 | 95 | no | | go |
| rain | 19 | 70 | yes | | dont_go |
| rain | 23 | 80 | yes | | dont_go |
| rain | 25 | 81 | no | | go |
| rain | 21 | 80 | no | | go |

Aprendizado de Máquina

Classificador



Avaliação

Processo de Classificação



Conhecimento de Fundo

Conhecimento de Fundo

Variáveis Independentes

Variável Dependente

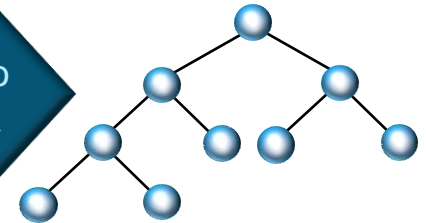
| | X ₁ | X ₂ | X ₃ | X ₄ | Y |
|----------|----------------|----------------|----------------|----------------|---------|
| sunny | 25 | 72 | yes | | go |
| sunny | 28 | 91 | yes | | dont_go |
| sunny | 22 | 70 | no | | go |
| sunny | 23 | 95 | no | | dont_go |
| sunny | 30 | 85 | no | | dont_go |
| overcast | 23 | 90 | yes | | go |
| overcast | 29 | 78 | no | | go |
| overcast | 19 | 65 | yes | | go |
| overcast | 26 | 75 | no | | go |
| overcast | 20 | 87 | yes | | dont_go |
| rain | 22 | 95 | no | | go |
| rain | 19 | 70 | yes | | dont_go |
| rain | 23 | 80 | yes | | dont_go |
| rain | 25 | 81 | no | | go |
| rain | 21 | 80 | no | | go |

Pode ser usado para fornecer informação já conhecida ao indutor

Especificação do Problema

Aprendizado de Máquina

Classificador



Avaliação

Pode ser usado ao selecionar os dados

Dados Brutos



Processo de Classificação



Especificação do Problema

Conhecimento de Fundo



Conhecimento de Fundo

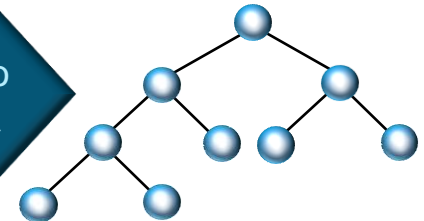
Variáveis Independentes

Variável Dependente

| | X ₁ | X ₂ | X ₃ | X ₄ | Y |
|--|----------------|----------------|----------------|----------------|---------|
| | sunny | 25 | 72 | yes | go |
| | sunny | 28 | 91 | yes | dont_go |
| | sunny | 22 | 70 | no | go |
| | sunny | 23 | 95 | no | dont_go |
| | sunny | 30 | 85 | no | dont_go |
| | overcast | 23 | 90 | yes | go |
| | overcast | 29 | 78 | no | go |
| | overcast | 19 | 65 | yes | go |
| | overcast | 26 | 75 | no | go |
| | overcast | 20 | 87 | yes | dont_go |
| | rain | 22 | 95 | no | go |
| | rain | 19 | 70 | yes | dont_go |
| | rain | 23 | 80 | yes | dont_go |
| | rain | 25 | 81 | no | go |
| | rain | 21 | 80 | no | go |

Aprendizado de Máquina

Classificador



O classificador gerado é avaliado e o processo pode ser repetido



Avaliação

Processo de Classificação



Especificação do Problema

Conhecimento de Fundo



Conhecimento de Fundo

Variáveis Independentes

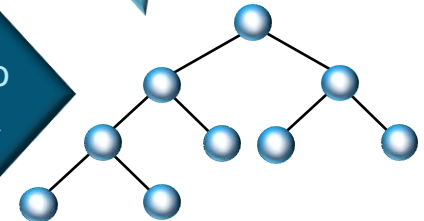
Variável Dependente

| | X ₁ | X ₂ | X ₃ | X ₄ | Y |
|----------|----------------|----------------|----------------|----------------|---------|
| sunny | 25 | 72 | yes | | go |
| sunny | 28 | 91 | yes | | dont_go |
| sunny | 22 | 70 | no | | go |
| sunny | 23 | 95 | no | | dont_go |
| sunny | 30 | 85 | no | | dont_go |
| overcast | 23 | 90 | yes | | go |
| overcast | 29 | 78 | no | | go |
| overcast | 19 | 65 | yes | | go |
| overcast | 26 | 75 | no | | go |
| overcast | 20 | 87 | yes | | dont_go |
| rain | 22 | 95 | no | | go |
| rain | 19 | 70 | yes | | dont_go |
| rain | 23 | 80 | yes | | dont_go |
| rain | 25 | 81 | no | | go |
| rain | 21 | 80 | no | | go |

Aprendizado de Máquina

Classificador deve fornecer uma descrição compacta do conceito existente nos dados

Classificador



Avaliação

Classificador pode ser uma caixa preta...



Especificação do Problema

Conhecimento de Fundo



Conhecimento de Fundo

Variáveis Independentes

Variável Dependente

| | X ₁ | X ₂ | X ₃ | X ₄ | Y |
|----------|----------------|----------------|----------------|----------------|---------|
| sunny | 25 | 72 | yes | | go |
| sunny | 28 | 91 | yes | | dont_go |
| sunny | 22 | 70 | no | | go |
| sunny | 23 | 95 | no | | dont_go |
| sunny | 30 | 85 | no | | dont_go |
| overcast | 23 | 90 | yes | | go |
| overcast | 29 | 78 | no | | go |
| overcast | 19 | 65 | yes | | go |
| overcast | 26 | 75 | no | | go |
| overcast | 20 | 87 | yes | | dont_go |
| rain | 22 | 95 | no | | go |
| rain | 19 | 70 | yes | | dont_go |
| rain | 23 | 80 | yes | | dont_go |
| rain | 25 | 81 | no | | go |
| rain | 21 | 80 | no | | go |

Aprendizado de Máquina

Classificador



Avaliação

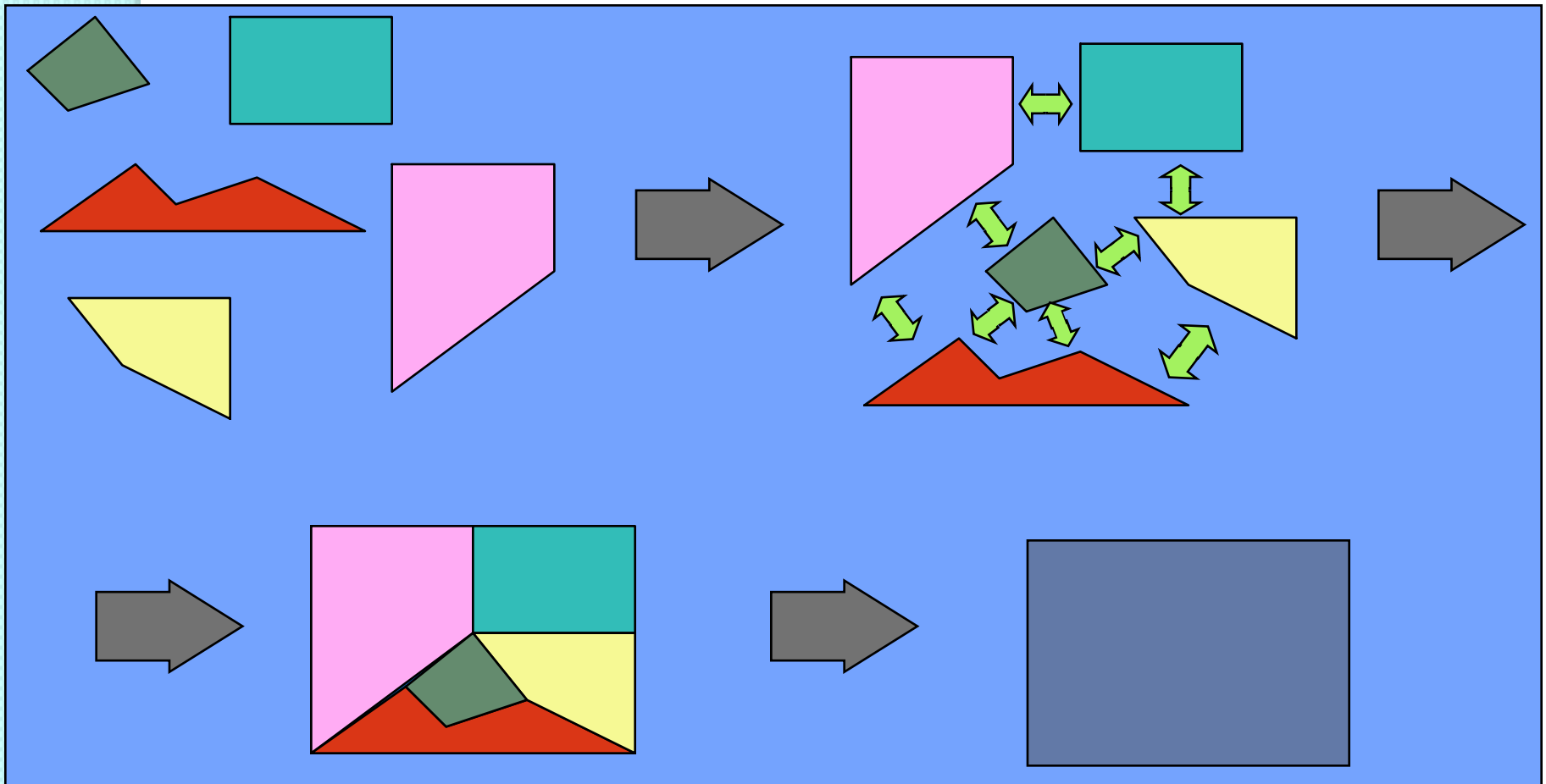
Linguagens de Descrição

- Tradução de problemas para uma linguagem de representação computacional
- São necessárias linguagens para descrever os exemplos, hipóteses induzidas e o conhecimento de fundo:
 - *Instance Description Language* (IDL)
 - *Hypotheses Description Language* (HDL)
 - *Background Knowledge Language* (BDL)

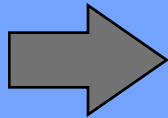
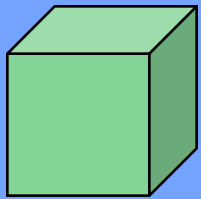
Linguagem de Descrição dos Exemplos

- *Instance Description Language (IDL)*
- 2 tipos:
 - Descrições estruturais
 - Descrições de atributos

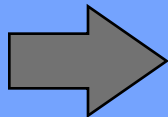
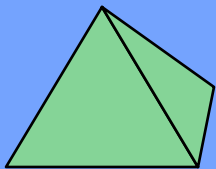
Descrições estruturais – um objeto é descrito em termos de seus componentes e a relação entre eles



Descrições de atributos – um objeto é descrito em termos de suas características globais como um vetor de valores de atributos



| Cubo | |
|------------------------|-------------------------|
| Número de faces | Polígono da face |
| 6 | quadrado |



| Pirâmide | |
|------------------------|-------------------------|
| Número de faces | Polígono da face |
| 4 | triângulo |

Linguagem de Descrição de Hipóteses e do Conhecimento de Fundo

- *Hypotheses Description Language (HDL)*
- *Background Knowledge Language (BDL)*

- Vários tipos:
 - Regras
 - Árvores de Decisão
 - Redes Semânticas
 - Lógica de Primeira Ordem
 - Funções Matemáticas

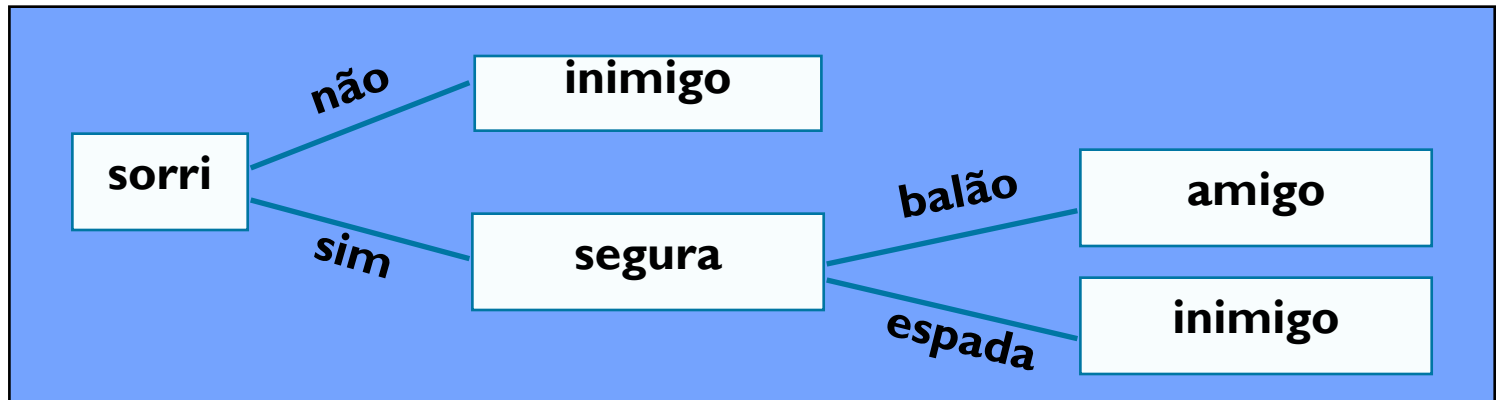
Formalismos usados em AM para descrever conceitos

- Regras se-então (if-then) para representar conceitos

Se Nublado *ou* Chovendo
então Levar_Guarda-Chuva

Formalismos usados em AM para descrever conceitos

- Árvores de decisão para representar conceitos



Linguagem de Descrição para alguns Indutores

| Indutor | IDL | HDL | BDL |
|------------|----------|--------------------------|--------------------------|
| C4.5 (J48) | Atributo | Atributo | |
| CN2 / PART | Atributo | Atributo | |
| Ripper | Atributo | Atributo | Atributo |
| FOIL | Atributo | Lógica de Primeira Ordem | Lógica de Primeira Ordem |
| RNA | Atributo | Função Matemática | |
| SVMs | Atributo | Função Matemática | |

Aprendizado Supervisionado

Cada exemplo é expresso por um conjunto de atributos



Aprendizado Supervisionado



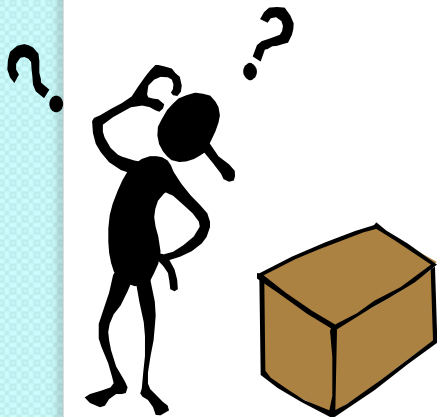
Objetiva moldar a estrutura de classificação para um problema específico, encontrando uma forma genérica de relatar um conceito.

Aprendizado Supervisionado



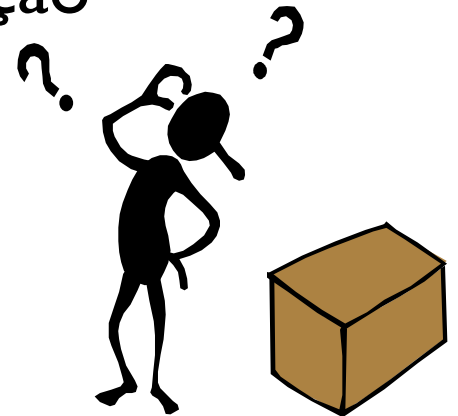
Categorias de Sistemas de Aprendizado

- Sistemas de aprendizagem são classificados em 2 categorias:
 - Sistemas “caixa-preta”
 - Sistemas “knowledge-oriented”



Categorias de Sistemas de Aprendizizado

- Sistemas “caixa-preta”:
 - Desenvolvem sua própria representação de conceitos
 - Representação interna não é facilmente interpretada por humanos
 - Não fornece esclarecimento ou explicação sobre o processo de classificação
 - Ex: RNAs



Categorias de Sistemas de Aprendizado

- Sistemas “orientados ao conhecimento”:
 - Criam estruturas simbólicas que podem ser compreendidas por seres humanos
 - Exs: regras de decisão, árvore de decisão, etc

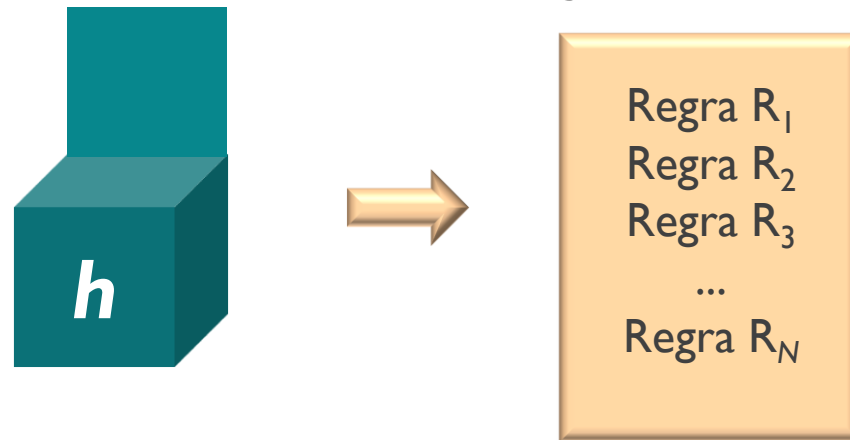


Conhecimento Adquirido (Hipótese h)

- h vista como classificador



- h vista como conjunto de regras



Categorias de Sistemas de Aprendizado

- A distinção entre essas duas categorias pode ser formulada em termos dos critérios:
 - **critério fraco:** o sistema usa dados para gerar subsídios para melhorar a performance com dados posteriores (ex: redes neurais, métodos estatísticos)
 - **critério forte:** o critério fraco é satisfeito e o sistema é capaz de comunicar sua representação interna na forma simbólica explicitamente
 - **critério ultra-forte:** os critérios fraco e forte são satisfeitos
 - O sistema deve ser capaz de comunicar sua representação interna na forma simbólica explicitamente **que pode ser usada por um humanos sem a ajuda de um computador**

Aprendizado Indutivo de Conceitos - AIC

- Dados $\varepsilon = \varepsilon^+ \cup \varepsilon^-$
 - Conjunto de exemplos de treinamento para aprendizado de um conceito C
- Encontrar uma hipótese h , expressa em uma linguagem de descrição L tal que:
 - Cada exemplo $e \in \varepsilon^+$ é coberto por h
 - Nenhum exemplo negativo $e \in \varepsilon^-$ é coberto por h

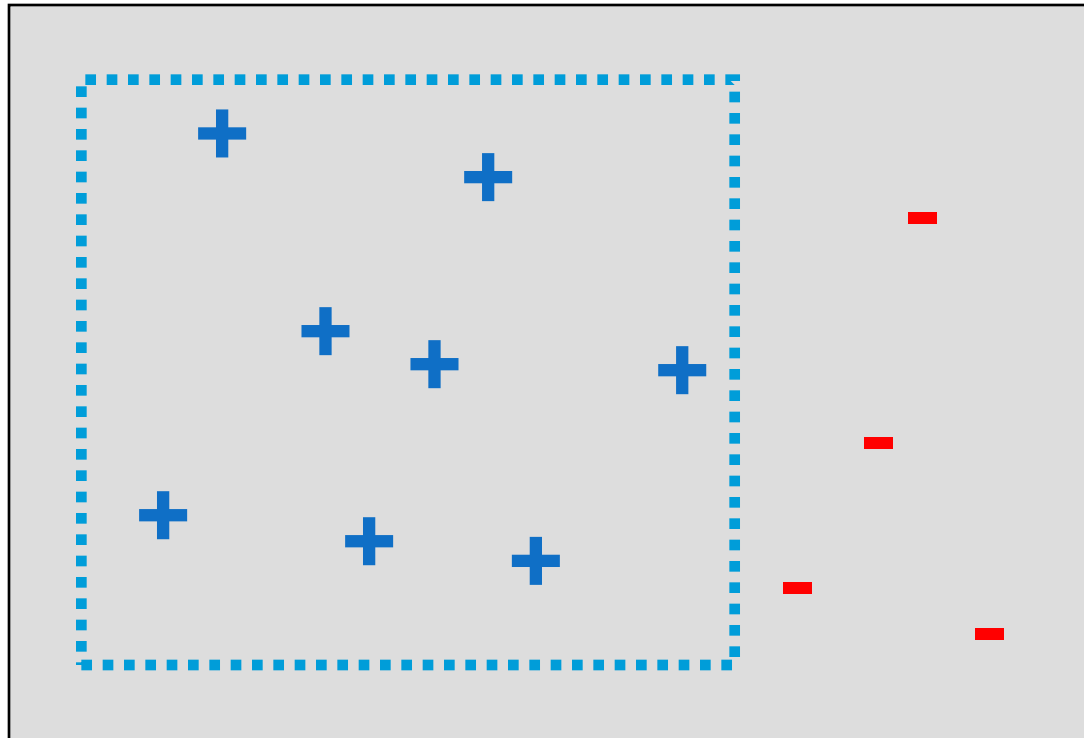
Aprendizado Indutivo de Conceitos - AIC (Cont)

$cobre(\mathbf{h}, \varepsilon) = \{e \in \varepsilon^+ \mid cobre(\mathbf{h}, e) = true\}$
(instância positiva)

$cobre(\mathbf{h}, \varepsilon) = \{e \in \varepsilon^- \mid cobre(\mathbf{h}, e) = false\}$
(instância negativa)

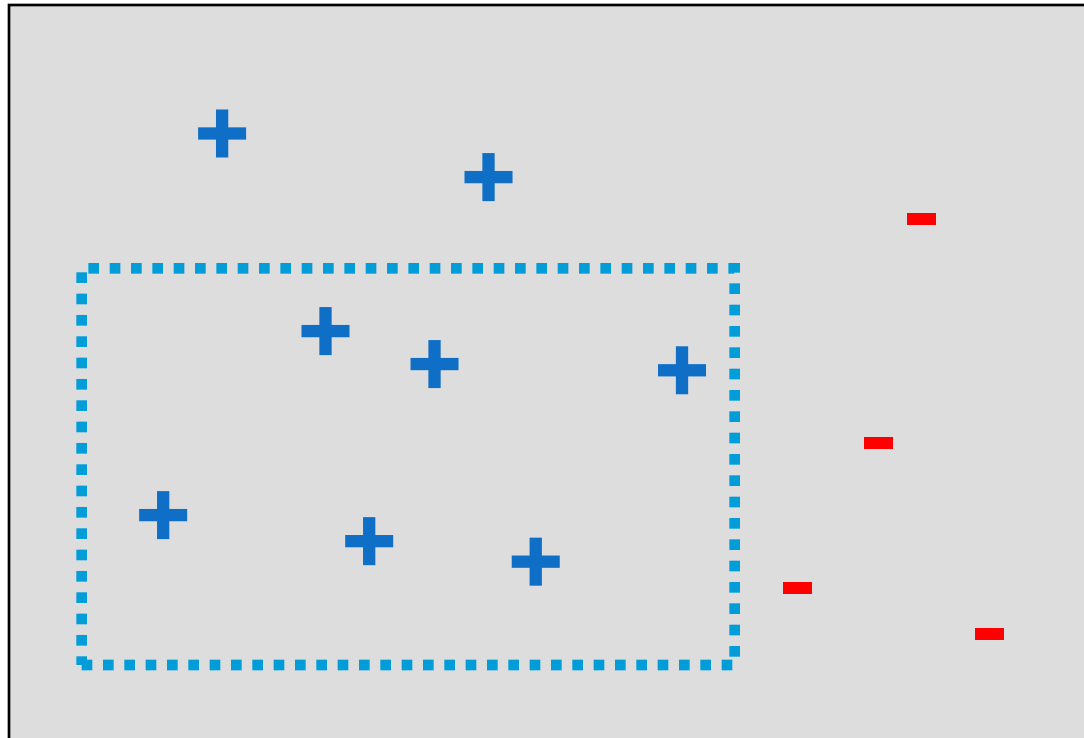
Consistência e Completeza de h

- h consistente e completa.
- h consistente e incompleta.
- h inconsistente e completa.
- h inconsistente e incompleta.



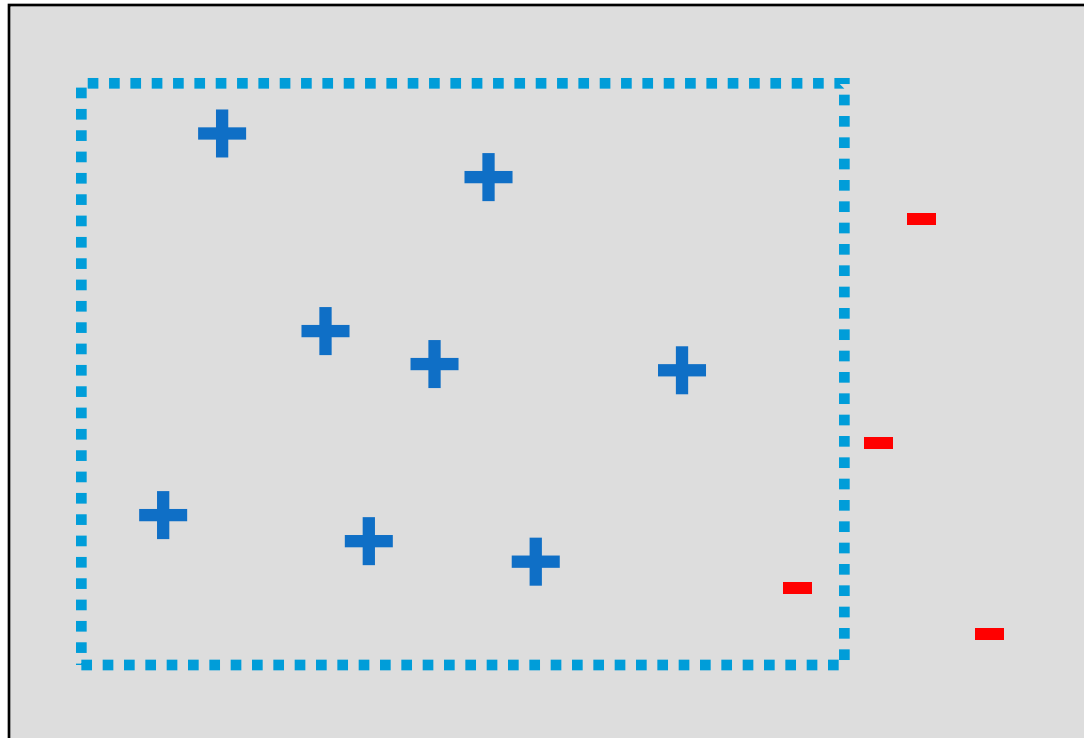
Consistência e Completeza de h

- h consistente e completa.
- h consistente e incompleta.
- h inconsistente e completa.
- h inconsistente e incompleta.



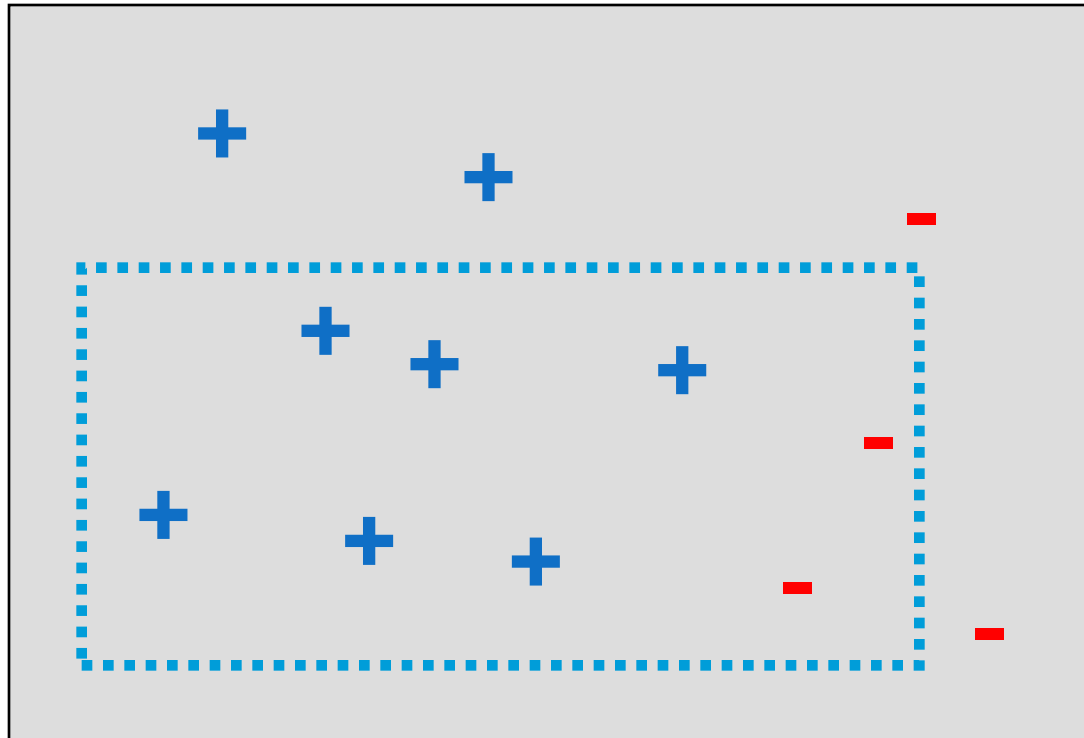
Consistência e Completeza de h

- h consistente e completa.
- h consistente e incompleta.
- h inconsistente e completa.
- h inconsistente e incompleta.

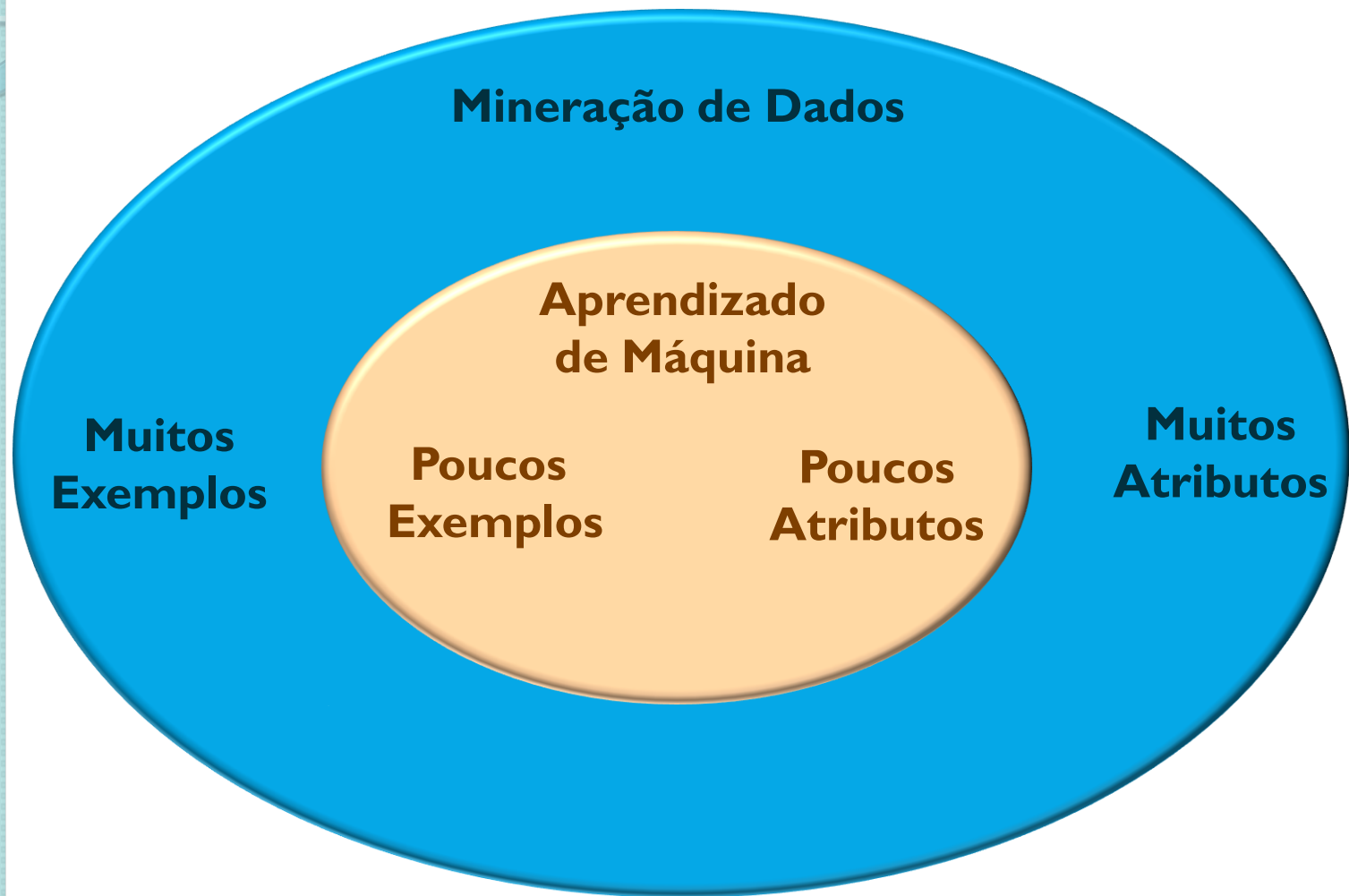


Consistência e Completeza de h

- h consistente e completa.
- h consistente e incompleta.
- h inconsistente e completa.
- h inconsistente e incompleta.



MD x AM



MD x AM



AM geralmente (mas não sempre) trabalha com pequena quantidade de dados, mas relevantes

MD x AM

Mineração de Dados

**Muitos
Exemplos**

**MD trabalha com grandes
bases de dados reais, sendo a
eficiência muito importante**

**Muitos
Atributos**

OLAP x MD

| Idade | Motivo | Duração | Valor | Risco |
|-------|--------|---------|--------|-------|
| 45 | Carro | 36 | 10,000 | Baixo |
| 20 | Negoc. | 20 | 35,000 | Alto |
| 37 | Casa | 40 | 30,000 | Baixo |
| 29 | Carro | 24 | 25,000 | Alto |
| 66 | Mobil. | 10 | 7,000 | Alto |

Se Idade ≥ 35 e Duração ≥ 20 então Risco = Baixo

WEKA – Software para apoiar a MD

- Weka – Waikato Environment for Knowledge Acquisition
 - Coleção de algoritmos de aprendizado de máquina e outras técnicas que dão suporte ao processo de MD
 - Implementação em Java
 - <http://www.cs.waikato.ac.nz/ml/weka/>
- Tutoriais (just a few):
 - <http://www.ibm.com/developerworks/br/opensource/library/os-weka1/>
 - http://forumsoftwarelivre.com.br/2011/arquivos/palestras/DataMining__Weka.pdf
 - <https://blog.itu.dk/SPVC-E2010/files/2010/11/wekatutorial.pdf>
- Curso online – Data Mining with Weka – Prof. Ian H. Witten
 - <http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/>