Inteligência Computacional

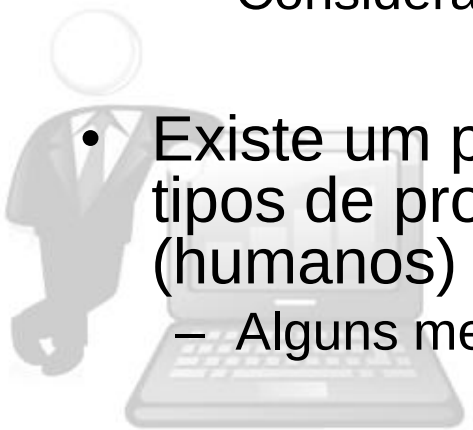
O que é Inteligência Computacional

Adaptado de notas de aula da Profa. Dra. Sarajane Marques Peres

Questões básicas

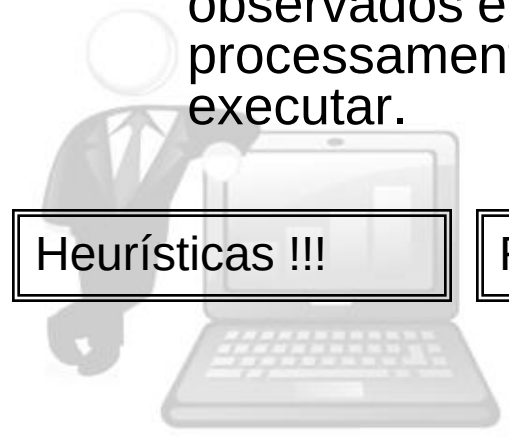
O que é Inteligência Artificial?

- É a ciência e engenharia de projetar/construir máquinas inteligentes, especialmente programas de computador inteligentes.
- Está relacionada ao uso de computadores para entender a inteligência humana (ou animal).
- Inteligência? É a parte computacional com habilidades para atingir metas no mundo (ambiente).
  - Considera graus de inteligência.
- Existe um problema em caracterizar, de forma geral, os tipos de procedimentos computacionais que nós (humanos) queremos chamar de inteligentes.
  - Alguns mecanismos são inteligentes, outros não.



Simulação da inteligência humana?

- Algumas vezes, mas nem sempre.
- Por um lado, nós podemos aprender alguma coisa sobre como fazer as máquinas resolverem problemas por meio da observação das pessoas ou apenas por meio do estudo de nossos próprios métodos.
- Por outro lado, a maioria dos estudos em IA envolve os problemas que o mundo apresenta e não o estudo de pessoas ou animais.
- A pesquisa em IA é livre para usar métodos que não são observados em pessoas ou que envolvem muito mais processamento computacional do que uma pessoa poderia executar.

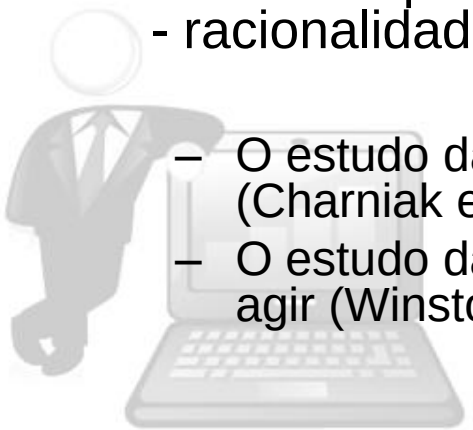


Heurísticas !!!	Representações de raciocínio !!!!	Computação Bio-Inspirada !!!

Definições para IA (Russell e Norvig, pg 5)

(Pensamento e raciocínio)

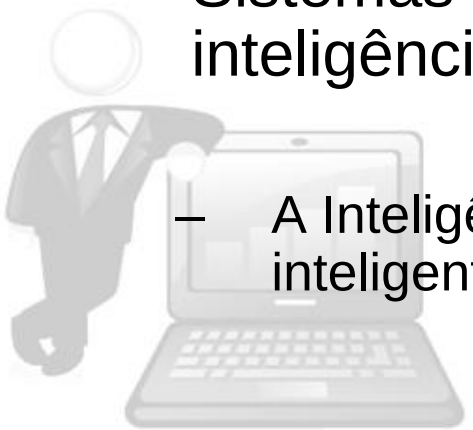
- Sistemas que pensam como seres humanos:
  - O novo e interessante esforço para fazer os computadores pensarem … máquinas com mentes, no sentido total e literal (Haugeland, 1985)
  - Automatização de atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado … (Bellman, 1978)
- Sistemas que pensam racionalmente (conceito ideal de inteligência - racionalidade)
  - O estudo das faculdades mentais pelo uso de modelos computacionais (Charniak e McDermott, 1985)
  - O estudo das computações que tornam possível perceber, raciocinar e agir (Winston, 1992)



Definições para IA (Russell e Norvig, pg 5)

(Comportamento)

- Sistemas que atuam como seres humanos
  - A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas (Kurzweil, 1990)
- Sistemas que atuam racionalmente (conceito ideal de inteligência - racionalidade)
  - A Inteligência Computacional é o estudo do projeto de agentes inteligentes (Poole et. al. 1998)



Com são feitos os estudos em IA?

- Os estudos em IA têm seu lado teórico e seu lado experimental. O lado experimental tem suas facetas básicas e aplicadas.

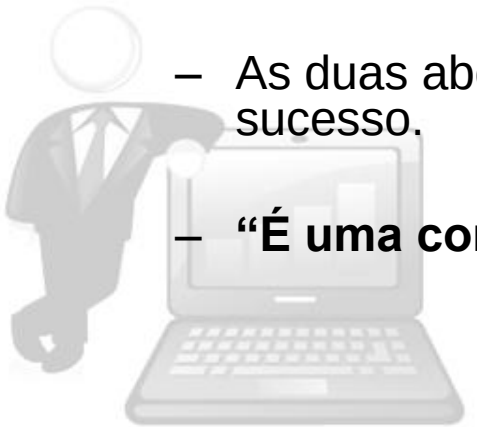
- Existem duas principais linhas de estudo:

  – Um é biológico, baseado em idéias que desde que os humanos são inteligentes, a IA deveria estudar humanos e imitar suas psicologia e fisiologia.

  – O outro é fenomenal, baseado em estudos e formalização de fatos de senso comum sobre o mundo e sobre os problemas que o mundo apresenta no alcance de objetivos.

  – As duas abordagem interagem em algum nível, e ambas alcançam o sucesso.

  – **“É uma corrida, mas ambos os corredores parecem estar andando.”**



Quando a pesquisa em IA começou?

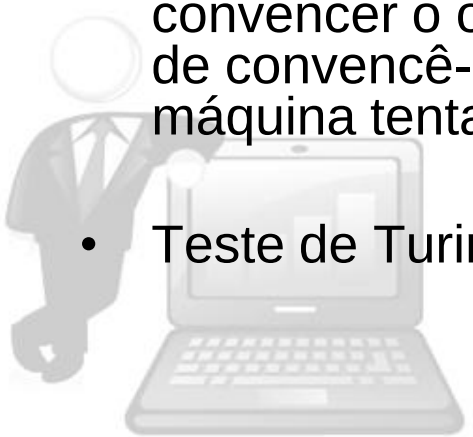
- Depois da segunda guerra mundial algumas pessoas, de forma independente, começaram a trabalhar em máquinas inteligentes.
- O matemático inglês Alan Turing parece ter sido o primeiro.
- Ele também foi o primeiro a dizer que a IA está mais ligada a programaçao de computadores do que a construção de máquinas.

Representação simbólica do conhecimento.
Inferência sobre representações do conhecimento.

Representação não simbólica do conhecimento.
Arquitetura conexionista.

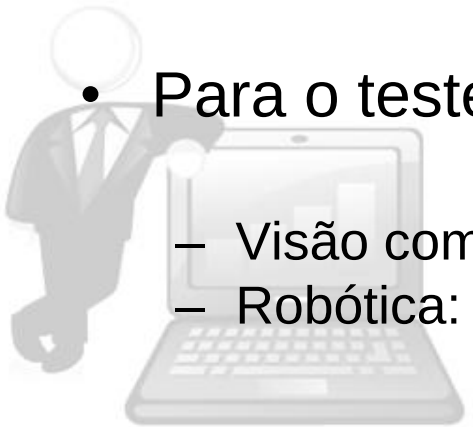
O que é o Teste de Turing?

- No artigo Computing Machinery and Intelligence, Turing discutiu as condições para considerar que uma máquina é inteligente.
- Ele argumentou que se a máquina pudesse se passar por um humano mediante um observador inteligente, então certamente ela poderia ser considerada inteligente. Este teste satisfaria a maioria das pessoas mas não todos os filósofos.
- Neste teste, o observador poderia interagir com a máquina e com um humano por meio de digitação, e então o humano tentaria convencer o observador (que era também um humano) no sentido de convencê-lo que se tratava realmente de um humano, e a máquina tentaria enganar o observador.
- Teste de Turing total: inclui um sinal de vídeo.



Como passar no teste?

- Para o teste tradicional:
  - Processamento de linguagem natural: para permitir que ele se comunique com sucesso em um idioma natural.
  - Representação do conhecimento: para armazenar o que sabe e "ouve".
  - Raciocínio automatizado: para usar as informações armazenadas com a finalidade de responder a perguntas e tirar novas conclusões
  - Aprendizado de máquina: para se adaptar a novas cincunstâncias e para detectar e extrapolar padrões
- Para o teste total:
  - Visão computacional: para perceber objetos
  - Robótica: para manipular objetos e movimentar-se



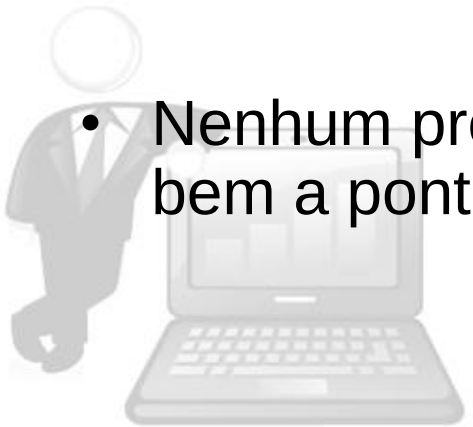
A IA está longe de alcançar o nível de inteligência humana?

- As poucas pessoas que acham que o nível de inteligência humana pode ser alcançado por meio da escrita de um grande número de programas do tipo “pessoa” estão agora escrevendo e montando vastas bases de conhecimento de fatos, usando linguagens de representação de conhecimento.

- Contudo, a maioria dos pesquisadores em IA acreditam que são necessárias idéias fundamentalmente novas, e portanto, não é possível predizer quando o nível de inteligência humana será alcançado.

E sobre “máquinas infantis” (que poderiam melhorar por meio de leitura e aprendizado)?

- Tal idéia foi proposta várias vezes, desde 1940.
- Eventualmente isso é trabalhado. Contudo, programas de IA não alcançaram ainda o nível de serem capazes de aprender muito mais do que uma criança aprende a partir de uma experiência física.
- Nenhum programa existente entende a linguagem tão bem a ponto de aprender por meio da leitura.



Um programa poderia elevar o seu nível de inteligência?

- Talvez sim.
- O problema é que a IA não está no nível de fazer com que tal processo “inicie”.

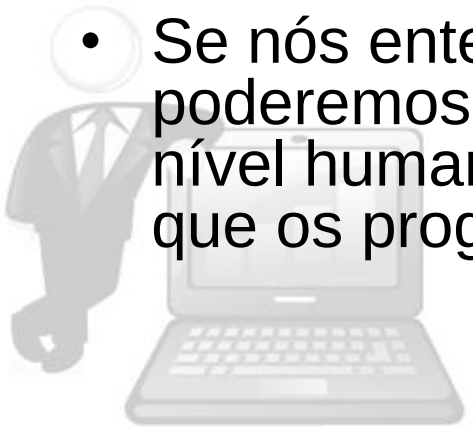
Qual é a relação entre IA e Filosofia?

- A IA tem muitas relações com filosofia, especialmente com a filosofia analítica moderna.
  - Ambas estudam a mente e ambas estudam o senso comum.

- Referencia:
  - Richmond Thomason. **Logic and artificial intelligence**. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2003. http://plato.stanford.edu/archives/fall2003/entries/logic-ai/.

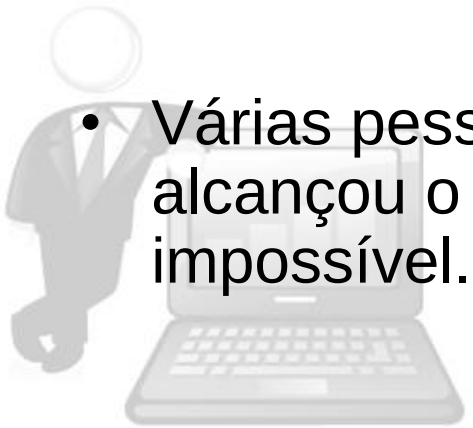
E o xadrez?

- Jogar xadrez requer certos mecanismos intelectuais e não outros.
- Programa de xadrez agora tem o papel de nível “grandmaster”, mas eles o fazem com mecanismos intelectuais limitados se comparados com aqueles usados pelos jogadores de xadrez humanos, que substituem uma grande quantia de computação por entendimento.
- Se nós entendermos melhor tal mecanismo, nós poderemos construir programas que joguem xadrez no nível humano, que fazem muito menos computação do que os programas atuais.



Algumas pessoas não dizem que a IA é uma idéia ruim?

- O filósofo John Searle diz que a idéia de uma máquina não biológica ser inteligente é incoerente.
- O filósofo Hubert Dreyfus diz que a IA é impossível.
- O cientista da computação Joseph Weizenbaum diz que a idéia é obscena, anti-humana e imoral.
- Várias pessoas tem dito que desde que a IA não alcançou o nível humano até agora, é porque deve ser impossível.



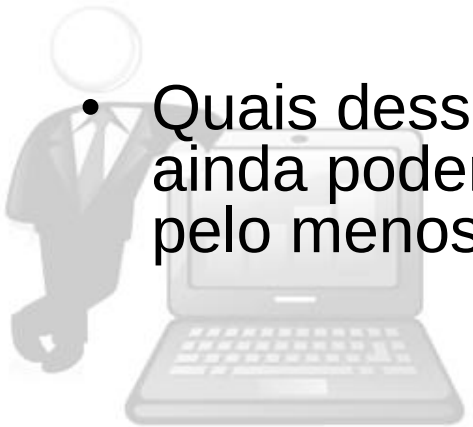
Argumento de inaptidão

Uma máquina nunca poderá fazer X

- Como exemplos de X, Turing listou:

  Ser amável, diligente, bonito, amigável, ter iniciativa, senso de humor, distinguir o certo do errado, cometer enganos, apaixonar-se, gostar de morangos e crème, fazer alguém se apaixonar por ela, aprende a partir da experiência, usar palavras corretamente, ser o sujeito de seu próprio pensamento, ter tanta diversidade de comportamento quanto o homem, fazer algo realmente novo.

- Quais dessas “inaptidões” foram alcançadas? Quais ainda podem ser? Quais não poderiam ser alcançadas, pelo menos com o estado atual da IA?



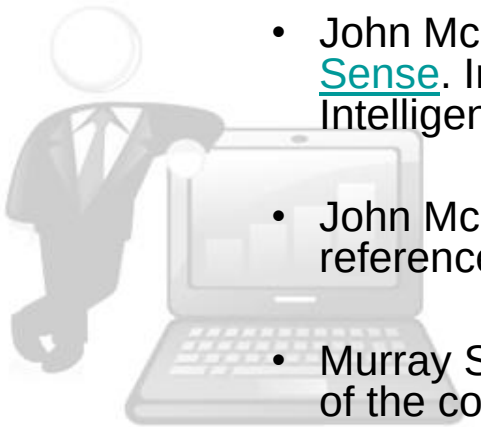
E quanto a teoria da computabilidade e complexidade computacional?

- Estas teorias são relevantes mas não fazem parte dos problemas **fundamentais** da IA .
- O que é importante para a IA é ter algoritmos tão capazes quanto as pessoas são na resolução de problemas.
- A identificação de subdomínios para os quais algoritmos existem é importante, mas muito dos resolverdores de problemas da IA não estão associados a subdomínios identificados.

Ramos da IA

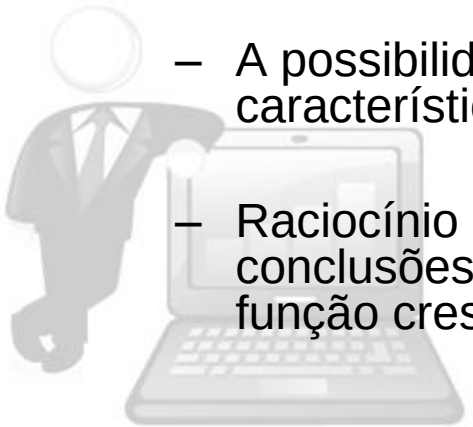
Ramos da IA

- **Inteligência Artificial Lógica**
  - O que um programa sabe sobre o mundo em geral, os fatos das situações específicas nas quais ele age, e suas metas, são representados por sentenças em alguma linguagem lógica matemática.
  - O programa decide o que fazer por inferir que certas ações são apropriadas para alcançar as metas.
  - Referências:
    - John McCarthy. Programs with Common Sense. In Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory, pages 77-84, London, U.K., 1959. Her Majesty's Stationery Office.
    - John McCarthy. Artificial Intelligence, Logic and Formalizing Common Sense. In Richmond Thomason, editor, Philosophical Logic and Artificial Intelligence. Klüver Academic, 1989.
    - John McCarthy. Concepts of Logical AI, 1996. Web only for now but may be referenced.
    - Murray Shanahan. Solving the Frame Problem, a mathematical investigation of the common sense law of inertia. M.I.T. Press, 1997.



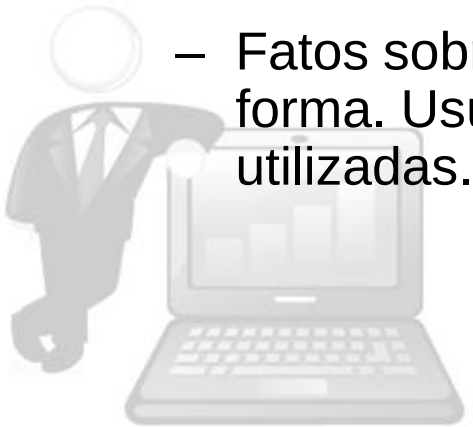
Ramos da IA

- **Inferência**
  - A partir de alguns fatos, outros podem ser inferidos.
  - Dedução lógica matemática é adequada para alguns propósitos, mas novos métodos de inferência não-monotônica têm sido adicionados para a lógica desde 1970.
  - O tipo mais simples de raciocínio não-monotônico é o raciocínio padrão no qual uma conclusão é inferida, mas pode ser retirada se existir uma evidência do contrário. Por exemplo:
    - Quando nós vemos um pássaro, nós inferimos que ele pode voar, mas esta conclusão pode ser refutada quando nós consideramos que é um pinguim.
  - A possibilidade que a conclusão tenha que ser retirada é que constitui a características não-monotônica do raciocínio.
  - Raciocínio lógico ordinário é monotônico desde que o conjunto de conclusões que pode ser projetado do conjunto de premissas é uma função crescente monotônica das premissas.



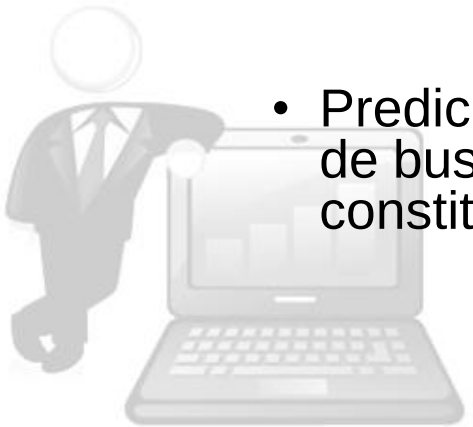
Ramos da IA

- Busca
  - Programas de IA frequentemente examinam um grande número de possibilidades, por exemplo, movimentos em um jogo de xadrez ou inferências por um programa de prova de teoremas.
  - Descobertas são continuamente feitas sobre como fazer isto de forma mais eficiente em diferentes domínios.
- Representação
  - Fatos sobre o mundo precisam ser representados de alguma forma. Usualmente as linguagens de lógica matemática são utilizadas.



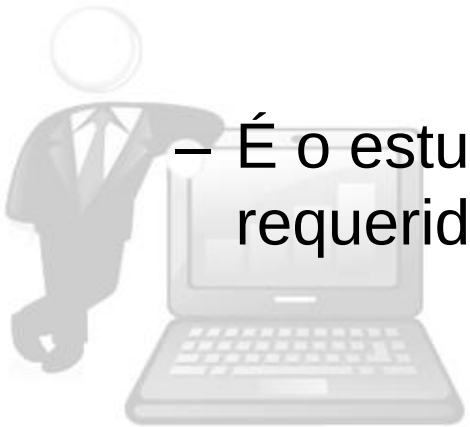
Ramos da IA

- Heurísticas
  - Uma heurística é uma maneira de tentar descobrir algo ou uma idéia incorporada a uma programa.
  - O termo é usado de várias formas na IA.
    - Funções heurísticas: são usadas em algumas abordagens para buscar uma medida do quanto um nó em uma árvore de busca está longe de uma meta.
    - Predicados heurísticos: comparam dois nós de uma árvore de busca para ver se um é melhor do que o outro, isto é, se constitui uma vantagem em direção à meta, ou se é mais útil.



Ramos da IA

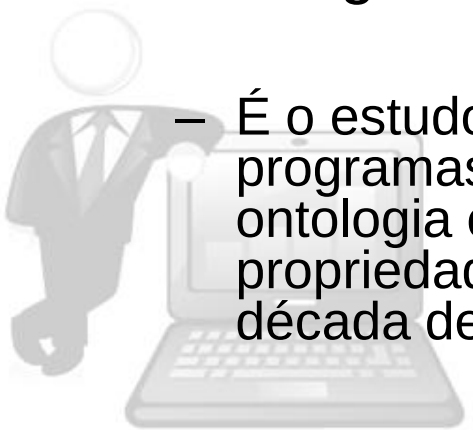
- **Raciocínio e conhecimento de senso comum**
  - Esta é a área na qual a IA está mais longe do nível humano, apesar do fato que tem sido uma área de pesquisa ativa desde a década de 50.
- Epistemologia
  - É o estudo dos tipos de conhecimento que são requeridos para resolver problemas no mundo



Ramos da IA

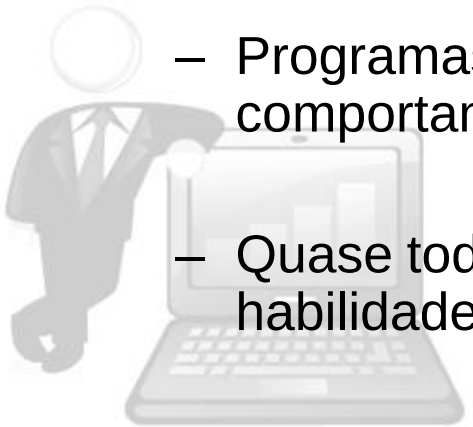
- **Planejamento**
  - Programas planejadores iniciam com fatos gerais sobre o mundo (especialmente fatos sobre os efeitos de ações), fatos sobre situações particulares e uma descrição de uma meta. A partir disto, eles geram uma estratégia para alcançar a meta. No mais comum dos casos, a estratégia é apenas uma sequência de ações.

- Ontologia
  - É o estudo dos tipos de coisas que existem. Em IA, os programas e sentenças lidam com vários tipos de objetos, e a ontologia estuda o que estes tipos são e quais são suas propriedades básicas. A ênfase na ontologia iniciou em na década de 90.



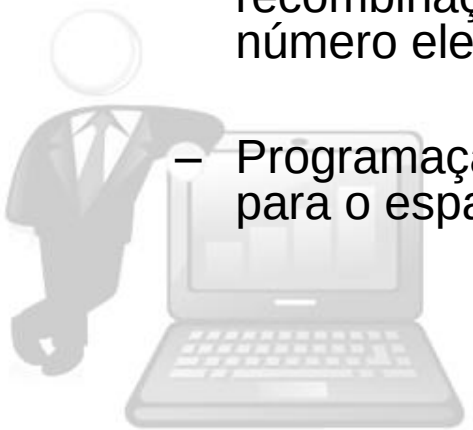
Ramos da IA

- **Aprendizado por experiência**
  - Programas que fazem algo.
  - As abordagens para a IA baseada em conexionismo ou redes neurais especializadas em alguma coisa.
  - Existem também o aprendizado de leis expressas em lógica.
  - Referência:
    - Tom Mitchell. Machine Learning. McGraw-Hill, 1997
  - Programas podem aprender apenas aqueles fatos ou comportamentos que seu formalismo pode representar.
  - Quase todos os sistemas de aprendizado são baseados nas habilidades limitadas da representação da informação.



Ramos da IA

- **Computação evolutiva**
  - Estratégias evolutivas: estratégias que otimizam parâmetros de valor real em sistemas dinâmicos;
  - Programação evolutiva: método em que os candidatos à solução de um dado problema são representados por máquinas de estados finitos, as quais evoluem pela mutação aleatória de seus diagramas de transição de estados, seguida pela seleção da mais bem adaptada;
  - Algoritmos genéticos: abstração da evolução biológica, tendo como inovações significativas a utilização conjunta de operadores de recombinação e inversão (além dos operadores de mutação) e de um número elevado de indivíduos em cada geração;
  - Programação genética: extensão das técnicas de algortimos genéticos para o espaço de programas computacionais



Ramos da IA

- **Reconhecimento de padrões**

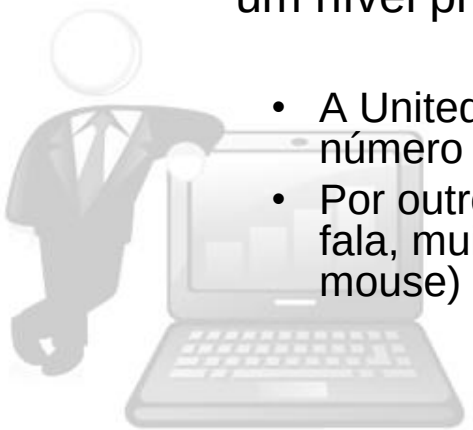
  – Quando um programa faz observações de algum tipo, ele está frequentemente programado para comparar o que ele “vê” naquele padrão.

  – Por exemplo: um programa de visão deve tentar “reconhecer” um padrão de olhos e de nariz em uma cena, com o objetivo de encontrar uma face.

Aplicações da IA

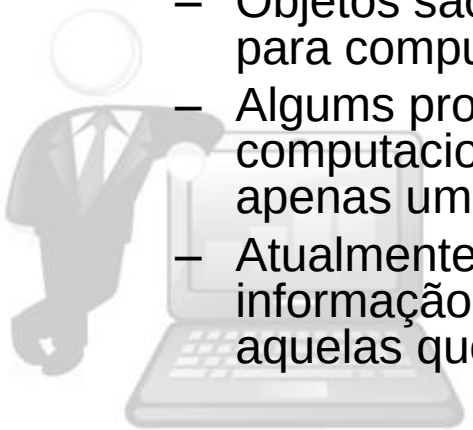
Aplicações da IA

- Jogos
  - Você pode comprar máquinas que podem jogar o mais alto nível de xadrez por algumas centena de dólares. Existe alguma Inteligência Artificial nela, mas elas jogam bem por conta, principalmente, da computação por força bruta – procurando centenas de milhares de posições. Para vencer o campeão do mundo por força bruta é necessário analisar 200 milhões de posições por segundo.
- Reconhecimento de Fala
  - Na década de 90, computadores reconhecedores de fala alcançaram um nível prático para propósitos limitados.
    - A United Airlines utiliza um sistema que usa reconhecimento de fala sobre número de vôos e nome de cidades. É muito conveniente.
    - Por outro lado, enquanto é possível instruir alguns computadores usando fala, muitos usuários tem voltado a usar sistemas convencionais (teclado e mouse) por parecerem ainda mais convenientes.



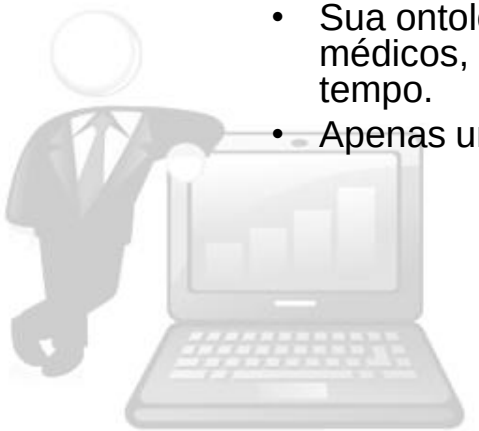
Aplicações da IA

- Entendimento de Linguagem Natural
  - Apenas colocar uma sequência de palavras dentro de um computador não é suficiente. Analisar (*parser*) sentenças também não é suficiente.
  - O computador tem que ser provido com um entendimento sobre o domínio do texto, e isto é atualmente possível apenas para domínios bastante limitados.
- Visão Computacional
  - Objetos são tri-dimensionais, mas as entradas para o olho humano, para computadores e câmeras de TV são bi-dimensionais.
  - Algums programas podem trabalhar em duas dimensões, mas visão computacional completa requer informação tri-dimensional que não é apenas um conjunto de visões bi-dimensionais.
  - Atualmente existem algumas maneiras limitadas de representação de informação tri-dimensional diretamente, e elas não são tão boas quanto aquelas que os humanos utilizam.



Aplicações da IA

- **Sistemas Especialistas**
  - Um “engenheiro do conhecimento” entrevista um especialista em um certo domínio e trata de incorporar seu conhecimento em um programa de computador para capacitá-lo a realizar alguma tarefa.
  - A qualidade do desempenho do programa depende se o mecanismo intelectual requerido para a tarefa esta dentro do que a IA é capaz de fazer atualmente.
  - Um dos primeiros sistemas especialistas foi o MYCIN (1974), o qual diagnosticava infecções bacterianas do sangue e sugeria tratamentos. Ele se saiu melhor do que estudantes de medicina e médicos profissionais, mas algumas limitações foram observadas.
    - Sua ontologia incluía bactérias, sintomas e tratamentos e não incluía pacientes, médicos, hospitais, dados sobre morte, recuperação ou outros eventos dependentes do tempo.
    - Apenas um paciente era considerado.



Aplicações da IA

- **Classificação heurística**

  – Um dos tipos mais úteis de sistemas especialistas é aquele capaz de colocar alguma informação dentro de um conjunto fixo de catergorias usando algumas fontes de informação.

  – Um exemplo é o sistema que aconselha sobre a aceitação de uma proposta de compra de cartão de crédito. Está disponível a informação sobre o proprietário do cartão de crédito, seus registros de pagamento e também sobre o item que ele está comprando e sobre o estabelecimento do qual ele está comprando (por exemplo, se existem dados sobre fraudes de cartão de crédito naquele estabelecimento).

- Baseado em:
  - WHAT IS ARTIFICIAL INTELLIGENCE? By John McCarthy. Stanford University - http://www-formal.stanford.edu/jmc/whatisai/
  - Russell e Norvig – Inteligência Artificial (livro texto) – Capítulo 1 e Capítulo 26.
  - Von Zuben - Notas de aula – Computação Evolutiva – FEEC/Unicamp.

Representação do Conhecimento

Gerenciando de conteúdo.

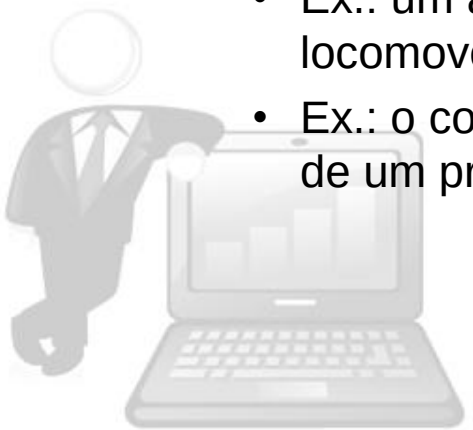
Como representar o conhecimento sobre fatos e sobre o mundo em estruturas passíveis de interpretação e processamento automático.

Usando simbolismo

Engenharia ontológica

- Domínios em “miniatura”:
  - Facilidade de criação de um vocabulário que o represente.
    - Ex.: um ambiente estático e simplificado onde um robô deve se locomover (a sala de uma máquina servidora de uma rede de computadores, por exemplo).
    - Ex.: o conhecimento para resolver um problema como parafusar uma placa de metal em uma estrutura padrão.

- Domínios complexos
  - Exigem representações gerais e flexíveis.
    - Ex.: um ambiente dinâmico e rico em informações onde um robô deve se locomover (os corredores de um hospital – pronto socorro, por exemplo).
    - Ex.: o conhecimento para resolver o problema de decidir qual é a causa de um problema que eventualmente ocorre em uma linha de produção.



Engenharia Ontológica

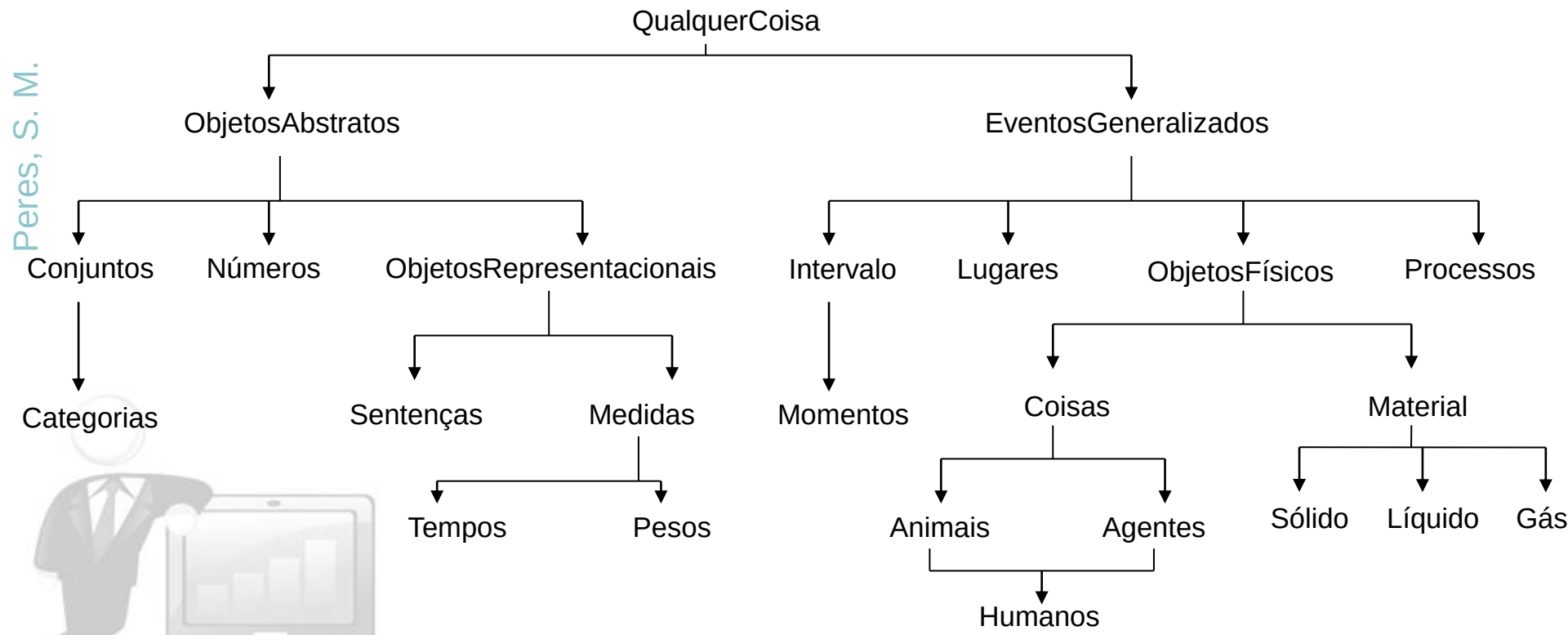
- Área de estudos referentes a elaboração de formas de representação de conceitos abstratos;
  - Ações, tempo, objetos físicos e crenças.

- Ontologia: (vocabulário sobre um domínio) – determina os tipos de itens que existem em um domínio;

- Ontologia superior: representa a estrutura geral de conceitos.
  - Define-se o que é um *objeto físico* mas deixa-se lacunas para serem preenchidas mais tarde, conforme necessidade, onde constarão informações detalhadas sobre diferentes objetos.

Engenharia Ontológica

- Um exemplo de ontologia superior. Cada arco indica que o conceito inferior é uma especialização do conceito superior



Engenharia Ontológica

- Ontologia de uso especial: possui um grande número de suposições simplificadoras.
  - Ex.: os números considerados são apenas aqueles pertencentes ao conjunto dos números inteiros.

- Ontologia de uso geral: a generalidade dos conceitos é *total.*
  - *Ex.: todos os números, racionais e irracionais, são considerados.*

- Para qualquer ontologia de uso especial, é possível fazer mudanças com a finalidade de obter uma generalidade maior.

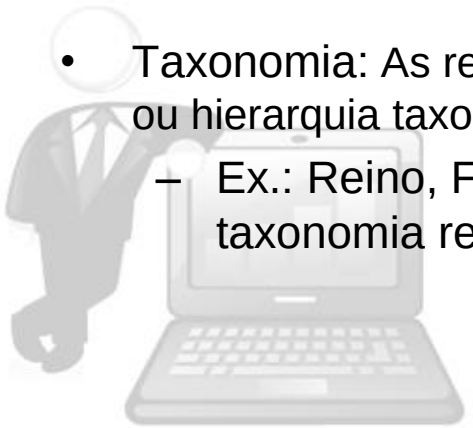
Categorias e objetos

- Um consumidor talvez tenha como objetivo comprar uma bola de basquete, e não uma determinada bola de basquete, como uma *Penalty.*
  - Bola de basquete: categoria
  - Bola de basquete Penalty: objeto

- Categorias permitem prognósticos: A partir de seu tamanho grande, da casca verde e rajada e da forma ovóide, pode-se deduzir que um objeto é uma melancia; a partir disso, é possível deduzir que ele seria útil em uma salada de frutas.

- Herança: Todas as instâncias da categoria Alimento são comestíveis e se Fruta é uma subclasse de Alimento e se Maçãs é uma subclasse de Fruta, então saberemos que toda maçã é comestível.

- Taxonomia: As relações de subclasses organizam as categorias em uma taxonomia ou hierarquia taxonômica.
  - Ex.: Reino, Filo, Classe, Ordem, Família, Gênero e Espécie ... formam uma taxonomia referente a organização dos seres vivos.



Categorias e objetos (representação)

- Um objeto é um elemento de (ou pertence a) uma categoria.

$$BolaPenalty \in BolasDeBasquete$$

- Uma categoria é uma subclasse de outra categoria.

$$BolasDeBasquete \subset Bolas$$

- Todos os elementos de uma categoria têm algumas propriedades.

$$x \in BolasDeBasquete \Rightarrow \mathrm{Re}\, donda(x)$$

- Os elementos de uma categoria podem ser reconhecidos por algumas propriedades.

$$Laranja(x) \wedge \mathrm{Re}\, donda(x) \wedge Diâmetro(x) = 23{,}75cm \wedge x \in Bolas \Rightarrow x \in BolasDeBasquete$$

- Uma categoria é um conjunto que tem algumas propriedades.

$$Cães \in EspéciesDomesticadas$$

Categorias e objetos

- Relações entre categorias (que não são subclasses)
  - Categorias disjuntas:

Disjuntos({Animais, Vegetais})

$$Disjuntos\ (s) \Leftrightarrow \left(\forall c_1, c_2 c_1 \in s \wedge c_2 \in s \wedge c_1 \neq c_2 \Rightarrow Interseção\left(c_1, c_2\right) = \{\ \}\right)$$

  - Decomposição exaustiva:

DecomposiçãoExaustiva({americanos, canadenses, mexicanos}, NorteAmericanos)

$$DecomposiçãoExaustiva\ (s, c) \Leftrightarrow \left(\forall i, i \in c \Leftrightarrow \exists\, c_2\ c_2 \in s \wedge i \in c_2\right)$$

Categorias e objetos

– Partição

Partição({Machos, Fêmeas},Animais)

$$Partição(s,c) \Leftrightarrow Disjuntos(s) \wedge DecomposiçãoExaustiva(s,c)$$

- Definição de uma categoria – condições necessárias e suficientes para pertinência

$$x \in Solteiros \Leftrightarrow NãoCasados(x) \wedge x \in Adultos \wedge x \in Machos$$

Composição Física

- Um objeto pode fazer parte de outro.
  - Relação *ParteDe*

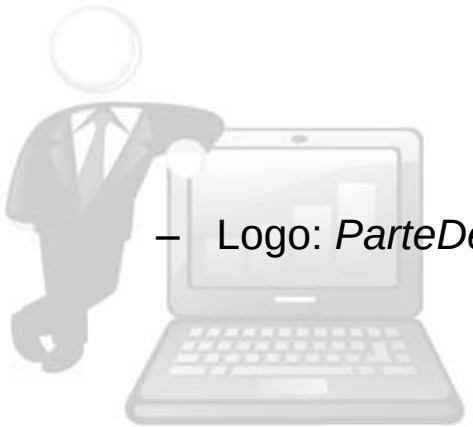
    Partede(Bucareste, Romênia)
    ParteDe(Romênia, EuropaOriental)
    ParteDe(EuropaOriental, Europa)
    ParteDe(Europa, Terra)

  - A relação *ParteDe* é transitiva e reflexiva, ou seja:

$$ParteDe(x, y) \wedge ParteDe(y, z) \Rightarrow ParteDe(x, z)$$

$$ParteDe(x, x)$$

  - Logo: *ParteDe(Bucareste, Terra)*



Composição Física

- Objetos Compostos: são caracterizados por relações estruturais entre partes.

$$Bípede(a) \Rightarrow \exists\, l_1, l_2, b\ Perna(l_1) \wedge Perna(l_2) \wedge Corpo(b) \wedge ParteDe(l_1, a) \wedge ParteDe(l_2, a) \wedge$$
$$ParteDe(b, a) \wedge \mathrm{Pr}esa(l_1, b) \wedge \mathrm{Pr}esa(l_2, b) \wedge l_i \neq l_2 \wedge$$
$$\left[\forall\, l_3\ Perna(l_3) \wedge ParteDe(l_3, a) \Rightarrow (l_3 = l_1 \vee l_3 = l_2)\right].$$

* Representação do conceito de “exatamente 2”

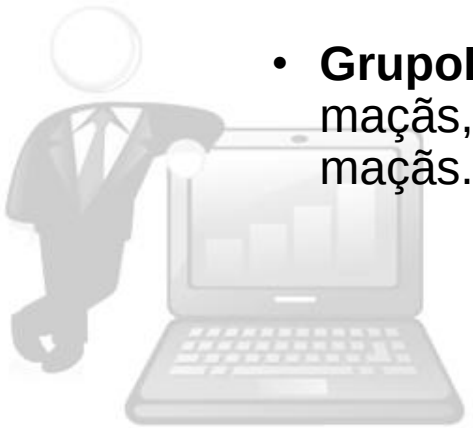
Composição Física

- Objetos compostos com partes definidas mas sem estrutura específica:
  - "As maçãs deste pacote pesam 900 gramas"
    - Não podemos atribuir o peso ao *conjunto* de maçãs, pois conjunto é um conceito matemático abstrato que tem elementos mas não tem *peso.*
    - **Grupo:** denota um objeto composto que consiste de outros objetos como partes (não como elementos)

      GrupoDe({Maçã1, Maçã2, Maçã3})

    - **GrupoDe(Maçãs)** é um objeto composto que consiste em todas as maçãs, não é a categoria nem tão pouco o conjunto de todas as maçãs.

      $$\forall x \; x \in s \Rightarrow ParteDe(x, GrupoDe(s)).$$



Medições

- Medidas: são valores atribuídos a propriedade de objetos.

Comprimento(L_1) = Polegadas(1,5) = Centímetros(3,81).

Centímetros(2,54 X d) = Polegadas(d).

Função de Unidades

- Outro exemplo:

$$d \in Dias \Rightarrow Duração(d) = Horas(24)$$

Medições

- Medidas não quantitativas;
  - Os exercícios têm dificuldade, sobremesas são deliciosas e os poemas têm beleza.
  - O aspecto mais importante destas medidas não é o dos valores numéricos específicos, mas sim o fato de que as medidas podem ser *ordenadas*.

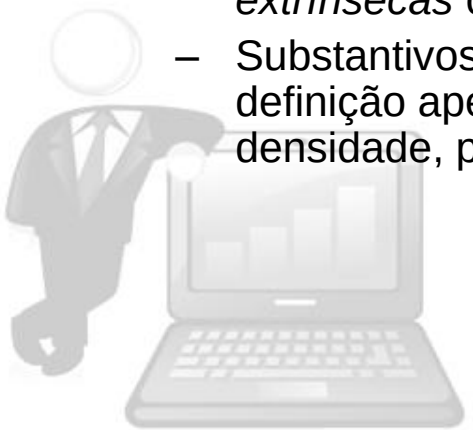
$$e_1 \in Exercícios \wedge e_2 \in Exercícios \wedge Escreveu(Norvig, e_1) \wedge Escreveu(Russel, e_2) \Rightarrow Dificuldade(e_1) > Dificuldade(e_2).$$

$$e_1 \in Exercícios \wedge e_2 \in Exercícios \wedge Dificuldade(e_1) > Dificuldade(e_2) \Rightarrow PontuaçãoEsperadada(e_1) < PontuaçãoEsperada(e_2)$$

Substâncias e Objetos

- **Material**: parte significativa da realidade que parece desafiar qualquer ***individuação*** óbvia – a divisão em objetos distintos.
  - Suponha que tenhamos um pouco de manteiga e um esquilo na nossa frente. Podemos dizer que existe um esquilo, mas não existe nenhum número óbvio de “objetos manteiga”, porque qualquer parte de um objeto manteiga também é um objeto manteiga (pelo menos até chegarmos a partes muito pequenas)
- A lingüística faz distinção entre *substantivos contáveis* e *substantivos de massa*.
  - Substantivos contáveis (COISA) designam objetos que incluem propriedades *extrínsecas* como peso, forma, comprimento …
  - Substantivos de massa (MATERIAL) designam objetos que incluem em sua definição apenas propriedades *intrínsecas* (pertinentes à substância do objeto – densidade, ponto de ebulição, sabor …)

$$x \in Manteiga \wedge ParteDe(y, x) \Rightarrow y \in Manteiga$$



Ações, Situações e Eventos

- Raciocínio sobre os resultados de ações:
  - como representar o modo como as ações afetam o mundo?

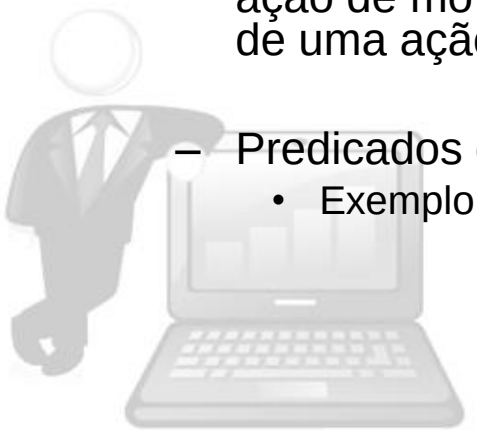
- Ontologia do **cálculo situacional**

  - Ações: são termos lógicos como *Avançar* ou *Virar(Direita)*.

  - Situações: são termos lógicos que consistem na situação inicial (S_0) e todas as situações que são geradas pela aplicação de uma ação a uma situação.
    - A função Resultado(a,s) identifica a **situação** que resulta quando a ação **a** é executada na situação **s** (veja Figura 10.2 (pág. 318))

  - Fluentes: são funções e predicados que variam de uma situação até a próxima (um agente pode deixar de segurar um objeto no curso de uma ação de movimento, a idade de um agente pode variar durante a execução de uma ação)

  - Predicados ou funções atemporais ou eternos.
    - Exemplo: PernaEsquerda(Agente) → será sempre a mesma enquanto ele existir.



Ações, Situações e Eventos

- Raciocínio sobre sequência de ações:
  - Execução de uma sequência vazia deixa a situação inalterada:

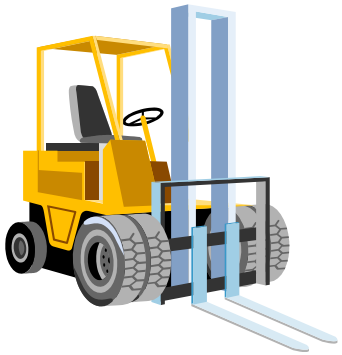
$$Resultado([\,],s) = s$$

  - Executar uma sequência não vazia é o mesmo que executar a primeira ação e depois executar o restante na situação resultante:

$$Resultado([a|seq], s) = Resultado(seq, Resultado(a,s))$$

Ações, Situações e Eventos

- Analisando um exemplo de uso:
  - Objetivo: um agente deve trazer o alvo para sua posição inicial.
  - Ações: ele pode IR de uma posição para uma posição adjacente.
  - Situação inicial:
    - o agente está em [1,1]
    - o alvo está em [1,2]
  - Os predicados fluentes:
    - Em (o,x,s)
    - Segurando (o,s)



Ações, Situações e Eventos

– A base de conhecimento inicial inclui a seguinte descrição:

Em (Agente, [1,1], S_0) e Em ($Alvo_1$, [1,2], S_0)

– Complementando a informação sobre o que não é verdadeiro em S_0.

$$Em(o, x, S_0) \Leftrightarrow [(o = Agente \wedge x = [1,1]) \vee (o = G_1 \wedge x = [1,2])].$$
$$\neg Segurando(o, S_0)$$

– Afirmando que G1 é o alvo e que as posições [1,1] e [1,2] são adjacentes

$$Alvo(G_1) \wedge Adjacente([1,1],[1,2]) \wedge Adjacente([1,2],[1,1]).$$

Ações, Situações e Eventos

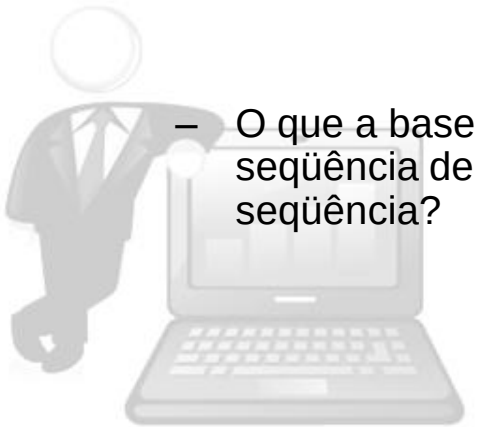
– Provando que o agente atinge seu objetivo indo até [1,2], agarrando o alvo e retornando a [1,1]. Isto é:

$$Em(G_1, [1,1], \text{Re}\,sultado([Ir([1,1],[1,2]), Agarrar(G_1), Ir([1,2],[1,1]), S_0)).$$

– É interessante a possibilidade de elaborar um plano para chegar ao alvo, o que é alcançado respondento à consulta “qual seqüência de ações resulta no fato de o alvo estar em [1,1]?”

$$\exists seq\ Em(G_1, [1,1], \text{Re}\,sultado(seq, S_0))$$

– O que a base de conhecimento deve conter para que seja possível estabelecer a seqüência de ações e provar que o agente atinge o seu objetivo se seguir tal seqüência?

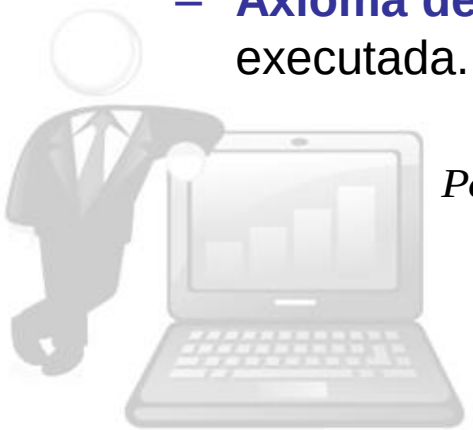


Ações, Situações e Eventos

- Cada ação é descrita por dois axiomas:
  - Assuma: *Poss(a,s)* para indicar que é possível executar a ação ***a*** na situação ***s***.
  - **Axioma da possibilidade**: afirma quando é possível executar a ação.

$$\text{Pr}econdições \Rightarrow Poss(a,s)$$

  - **Axioma de efeito**: afirma o que acontece quando uma ação possível é executada.



$$Poss(a,s) \Rightarrow Mudanças\ que\ resultam\ da\ execução da\ ação$$

Ações, Situações e Eventos

- Axiomas de possibilidade para o mundo em discussão afirmam que um agente pode seguir entre posições adjacentes, agarrar uma peça (alvo) na posição atual e soltar alguma peça (alvo) que esteja segurando.

$$Em(Agente, x, s) \wedge Adjacente(x, y) \Rightarrow Poss(Ir(x, y), s).$$

$$Alvo(g) \wedge Em(Agente, x, s) \wedge Em(g, x, s) \Rightarrow Poss(Agarrar(g), s).$$

$$Segurando(g, s) \Rightarrow Poss(Soltar(g), s).$$

Ações, Situações e Eventos

- Axiomas de efeito afirmam que, se uma ação for possível, certas propriedades (fluentes) serão válidas na situação que resultar da execução da ação.
- Ir de x para y resulta em estar em y, agarrar o alvo resulta em segurar o alvo, e soltar o alvo resulta em não segurá-lo.

$$Poss(Ir(x,y),s) \Rightarrow Em(Agente,y,\text{Re}sultado(Ir(x,y),s)).$$

$$Poss(Agarrar(g),s) \Rightarrow Segurando(g,\text{Re}sultado(Agarrar(g),s)).$$

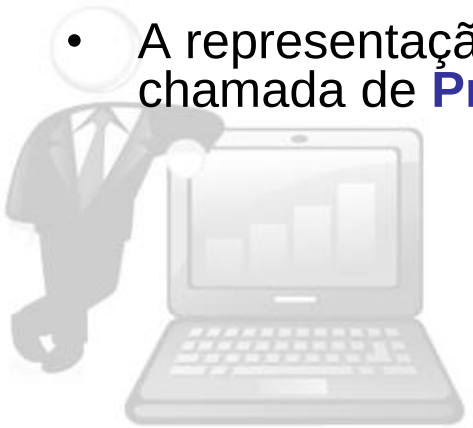
$$Poss(Soltar(g),s) \Rightarrow \neg Segurando(g,\text{Re}sultado(Soltar(g),s)).$$

Ações, Situações e Eventos

- Agora já podemos provar que nosso pequeno plano atinge a meta?
- Infelizmente não.
- Ir([1,1],[1,2]) é possível em S_0,então o agente alcança a posição [1,2].
- Intuitivamente, sabemos que o fato do agente ir até a posição [1,2] não modifica a posição do alvo. Mas não codificamos isso na nossa base de conhecimento pois:

Os axiomas de efeito afirmam o que muda, mas não o dizem o que permanece igual.

- A representação de todas as coisas que permanecem iguais é chamada de **Problema do Quadro.**



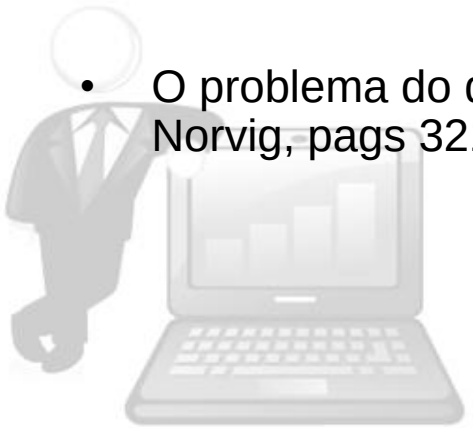
Ações, Situações e Eventos

- Assim, percebe-se que é necessário encontrar uma solução eficiente para o problema do quadro porque, no mundo real, quase tudo permanece igual durante quase todo o tempo. Cada ação afeta apenas uma minúscula fração de todos os fluentes.

- Uma abordagem é escrever axiomas do quadro explícitos que dizem o que permanece igual. Por exemplo, os movimentos do agente deixam outros objetos estacionários, a menos que eles estejam sendo segurados:

$$Em(o,x,s) \wedge (o \neq Agente) \wedge \neg Segurando(o,s) \Rightarrow Em(o,x,\text{Re}\,sultado\,(Ir(y,z),s)).$$

- O problema do quadro vai muito além deste exemplo. Veja em Russel e Norvig, pags 321, 322, 323 uma discussão sobre este problema.

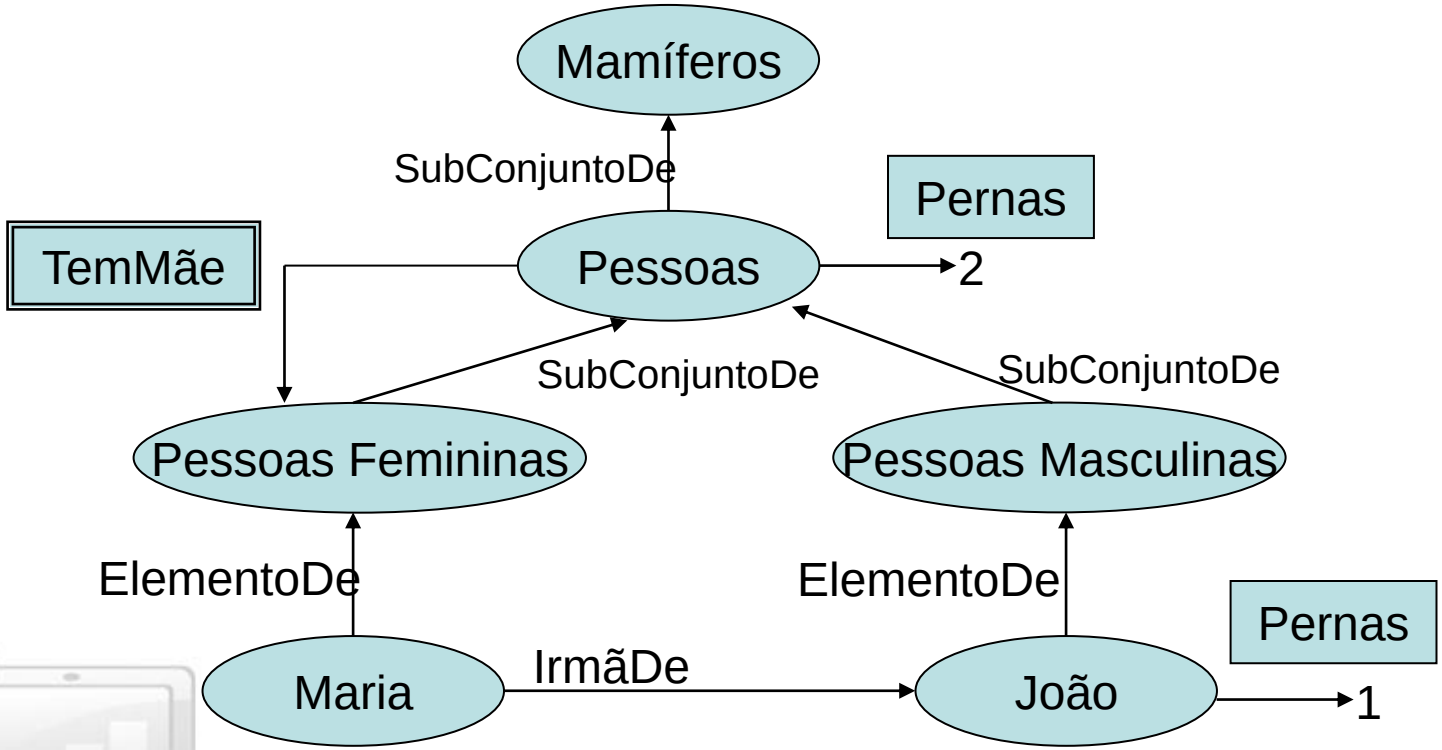


Redes Semânticas

- Sistema projetado para organizar e raciocinar com categorias.
  - Uma notação gráfica típica exibe nomes de objetos ou categorias de objetos em elipses ou retângulos e os conecta por meio de arcos rotulados.
  - O arco ElementoDe entre Maria e Pessoas Femininas corresponde à asserção lógica: *Maria* $\in$*PessoasFemininas*
  - O arco IrmãDe entre Maria e João corresponde à asserção lógica: *IrmãDe(Maria, João)*

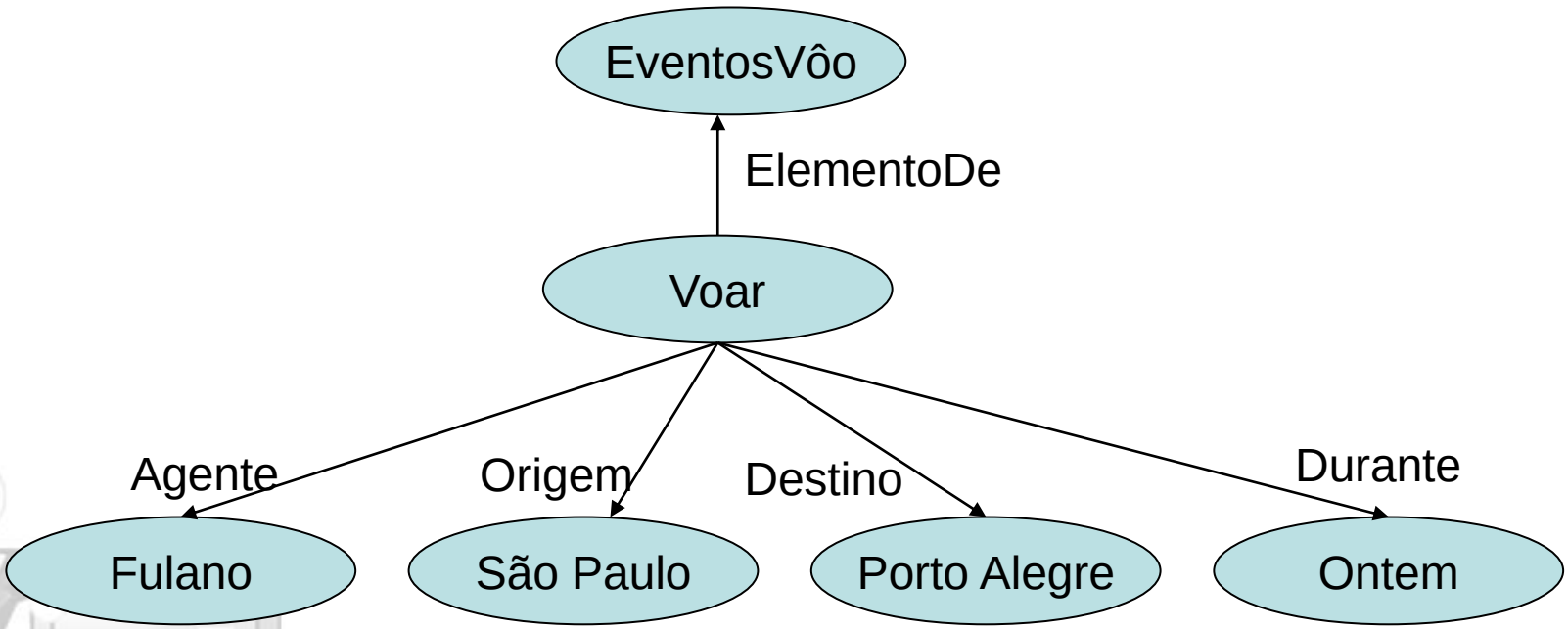
Redes Semânticas

- Um exemplo: uma rede semântica com quatro objetos (João, Maria, 1 e 2) e quatro categorias.



Redes Semânticas

- Outro exemplo: fragmento de uma rede semântica.



Outra forma de representação
(não simbólica)

Representação cromossômica

- Um cromossomo representa, para o algoritmo genético, uma possível solução para um problem.

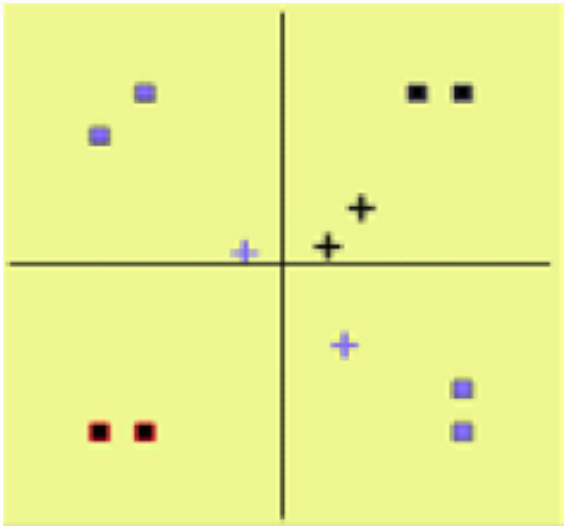
- Representação do problema das 8 rainhas
  - Cada posição do cromossomo pode ser valorada com números de 1 a 8, sendo que cada número representa a posição de cada rainha em uma coluna de um tabuleiro 8X8:

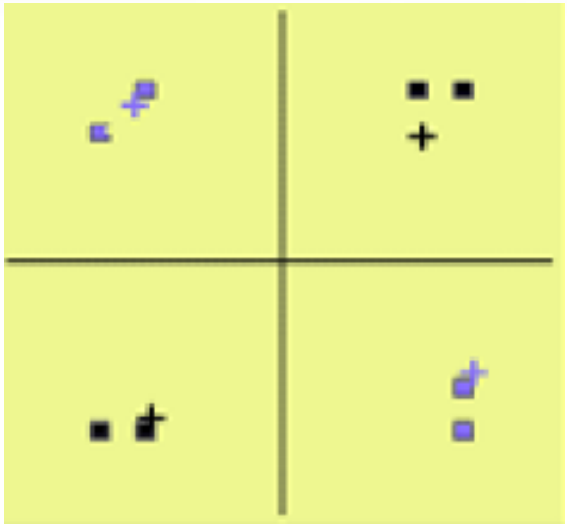
2 4 7 4 8 5 5 2
3 2 7 5 2 4 1 1

- Representação do problema do caixeiro viajante:
  - Cada posição do cromossomo representa o identificador de uma cidade e sua ordenação no cromossomo representa a ordem de visitação que o caixeiro deve realizar:

1 2 3 4 5 6 7 8 9 10 11 12
7 3 1 11 4 12 5 2 10 9 6 8

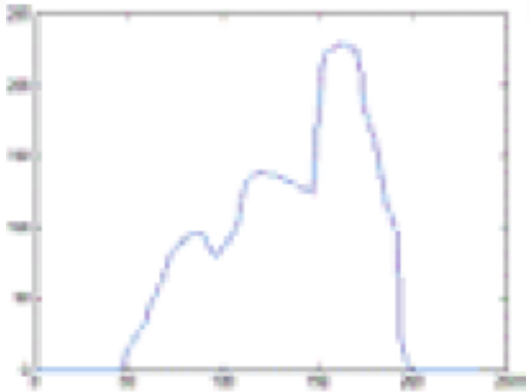
Representação neural do particionamento do espaço

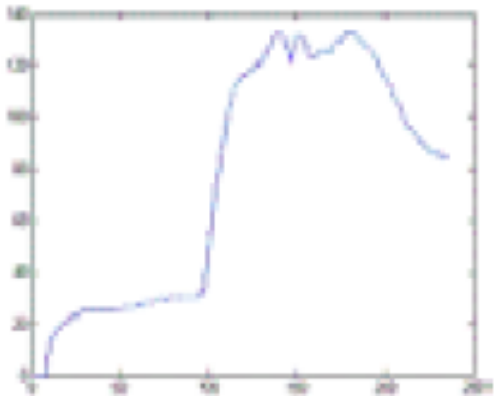




Pontos no R^2

Representação de uma imagem





(0 0 0 0 5 85 155 156 … 859 56 5 0 0 0)

Assinatura de bits (horizontal e vertical)

- Russel e Norvig
  - Capítulo 10 – Representação do conhecimento (pags. 307 – 321, 338 – 341, 115-116)
- LIBRAS Signals Recognition: a study with Learning Vector Quantization and Bit Signature (artigo científico, SBRN, 2006)
- Help Matlab – www.mathworks.com

Peres, S. M.

Descoberta de Conhecimento em Bases de Dados

Bibliografia

•Slides baseados nos Capítulos 1, 2 e 3 de:

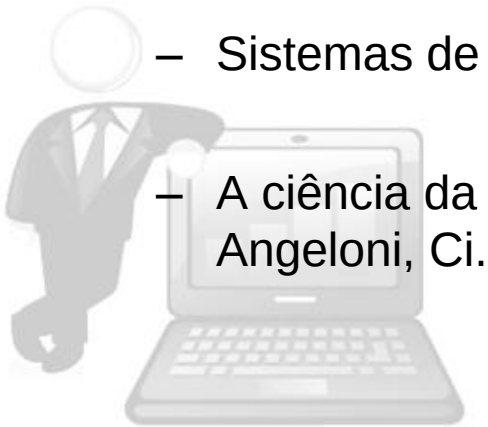
- Data Mining: Um guia prático. Goldschmidt e Passos. Editora Campus, 2005.

•Bibliografia Auxiliar

- From Data Mining to Knowledge Discovery: An Overview. Fayyad, Shapiro e Smith. *Knowledge Discovery and data Mining, Menlo Park: AAAI, 1996.*

- Sistemas de Banco de Dados. Elsmari e Navathe. (2004)

- A ciência da informação e a tomada de decisão, Maria Terezinha Angeloni, Ci. Inf., Brasília, v. 32, n. 1, p. 17-22, jan./abr. (2003).



Visão Geral

- Avanços na tecnologia – proliferação de bases de dados de diferentes naturezas.
  - Bases de dados da ordem de terabytes de informação.

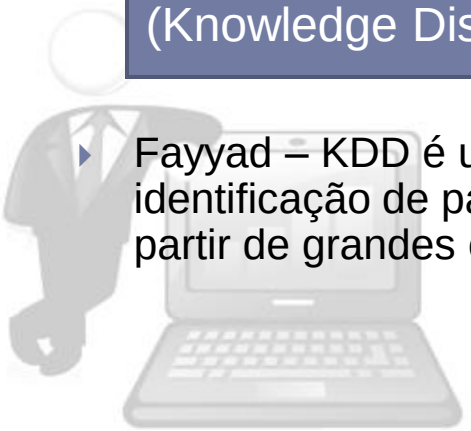
- Questões:
  - O que fazer com todos os dados armazenados?
  - Como utilizar o patrimônio digital em benefício das instituições?
  - Como analisar e utilizar de maneira útil todo o volume de dados disponível?

- Ferramentas para relacionar, analisar e interpretar esses dados, de forma automática e "inteligente", levando à concepção de estratégias de ação.

Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases - KDD)

Mineração de Dados (Data Mining)

- Fayyad – KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados.



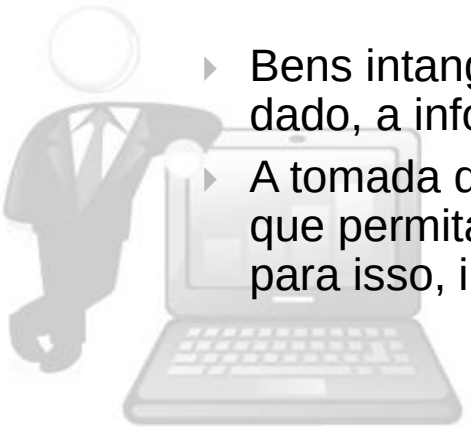
Peres, S. M.

Introdução

Dados – Informação - Conhecimento

- Dados:
  - Navathe: Os dados são fatos que podem ser gravados e que possuem um significado implícito.
  - Ex.:
    - nomes, endereços, saldos, preços, alturas, pesos, etc.
    - textos, videos, imagens, músicas, etc.
    - posicionamento geográfico, forma, etc.
    - atributos químicos, físicos, biológicos, etc.
  - A ciência da informação e a tomada de decisão:

(Ci. Inf., Brasília, v. 32, n. 1, p. 17-22, jan./abr. 2003 – Maria Terezinha Angeloni)



- Bens intangíveis importantes para o gerenciamento de uma organização: o dado, a informação e o conhecimento.
- A tomada de decisão necessita de um sistema de comunicação eficiente que permita a rápida circulação da informação e do conhecimento, sendo, para isso, indispensável o suporte da tecnologia.

Dados – Informação - Conhecimento

- Elementos intervenientes na tomada de decisão:

(Ci. Inf., Brasília, v. 32, n. 1, p. 17-22, jan./abr. 2003 – Maria Terezinha Angeloni)

- dado, informação,conhecimento, comunicação e tecnologia.

Motivando

Dado, informação e conhecimento são elementos que formam um sistema hierárquico de difícil delimitação.

O que é dado para um indivíduo pode ser informação e/ou conhecimento para outro.

Dados:

• elementos brutos, sem significado, desvinculados da realidade.
• observações sobre o estado do mundo.
• símbolos e imagens que não dissipam nossas incertezas.
• matéria-prima da informação.

" Dados sem qualidade levam a informações e decisões da mesma natureza. "

Informação:

• dados processados e contextualizados.
• desprovida de significado e de pouco valor.
• a matéria-prima para se obter conhecimento.

Conhecimento

• é a informação mais valiosa (...) é valiosa precisamente porque alguém deu à informação um contexto, um significado, uma interpretação (...).
• pode então ser considerado como a informação processada pelos indivíduos. O valor agregado à informação depende dos conhecimentos anteriores desses indivíduos.

" Conhecer é um processo de compreender e internalizar as informações recebidas, possivelmente combinando-as de forma a gerar mais conhecimento."

Elementos intervenientes na tomada de decisão:

(Ci. Inf., Brasília, v. 32, n. 1, p. 17-22, jan./abr. 2003 – Maria Terezinha Angeloni)

Reflita!!!

Visão Geral

- Etapas operacionais do Processo de KDD (forma resumida).

Analista humano: orienta a execução do processo

Captação
Organização
Tratamento

Busca de conhecimento útil

Tratamento do conhecimento obtido

Definições

- Padrão: deve ser interpretado como um conhecimento representado segundo as normas sintáticas de alguma linguagem formal (Fayyad).
- Padrão Compreensível: refere-se a um padrão representado em alguma forma de representação do conhecimento que possa ser interpretada pelo homem.
- Padrão Válido: indica que o conhecimento deve ser verdadeiro e adequado ao contexto da aplicação de KDD.
- Padrão Novo: deve acrescentar novos conhecimentos aos conhecimentos existentes no contexto da aplicação de KDD.
- Conhecimento Útil: é aquele que pode ser aplicado de forma a proporcionar benefícios ao contexto da aplicação de KDD.

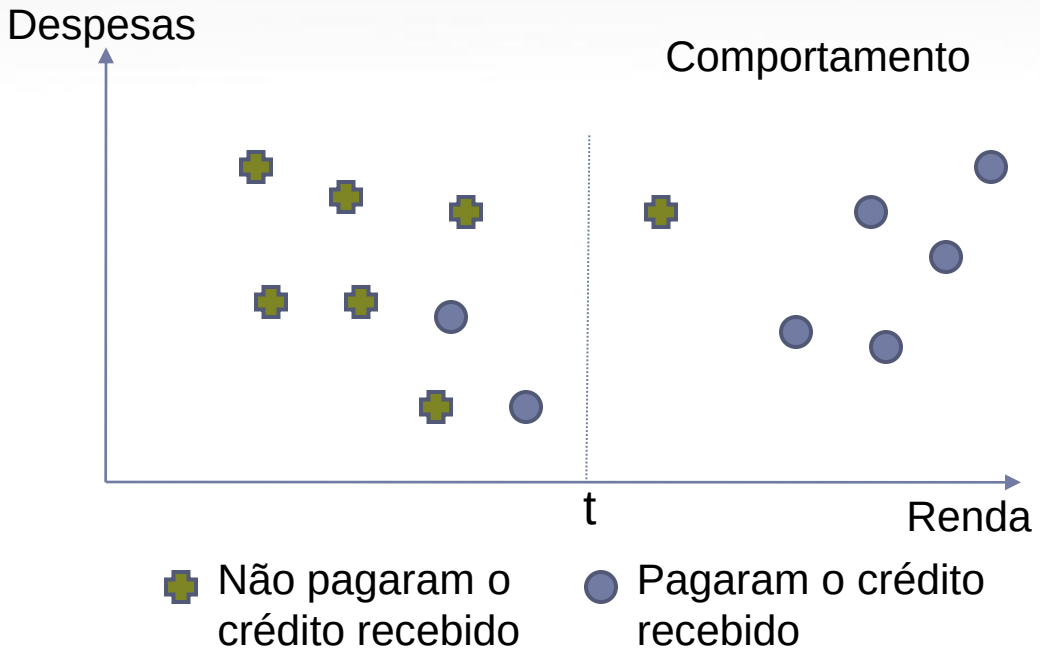
Visão Geral

- Base de fatos de uma financeira hipotética (Naliato)

SE Renda >= R$ t
ENTÃO Cliente = Não negligente
SENÃO Cliente = Negligente.

A regra é um padrão compreensível pelo homem.

Uma representação vetorial seria mais difícil de compreender.



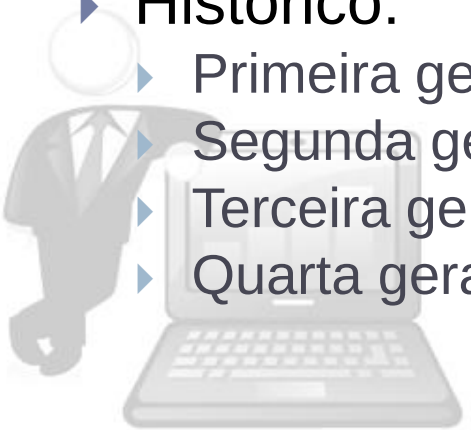
- Separabilidade linear: o problema acima não é linearmente separável.
- Acurácia (confiança) da regra: proporção dos casos que satisfazem ao antecedente e ao conseqüênte da regra em relação ao número de casos que satisfazem somente ao antecedente dessa regra.

Visão Geral

- Áreas envolvidas:
  - Estatística, Inteligência Artificial e Aprendizado de Máquina
    - Reconhecimento de Padrões
  - Banco de Dados

- Atividades;
  - Desenvolvimento Tecnológico
  - Execução de KDD
  - Aplicação de Resultados

- Histórico:
  - Primeira geração: utilização de algoritmos puros.
  - Segunda geração: utilização de *suites.*
  - Terceira geração: soluções orientadas a negócios.
  - Quarta geração: suporte ao processo de KDD.



Peres, S. M.

Conceitos Básicos

Caracterização

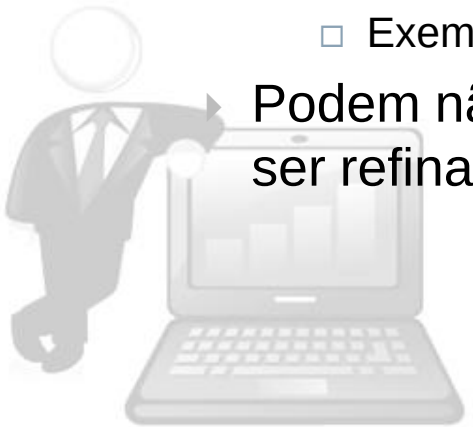
- Componentes de uma aplicação de KDD
  - O problema em que será aplicado o processo KDD.
  - Os recursos disponíveis para a solução do problema.
  - Os resultados obtidos a partir da aplicação dos recursos disponíveis em busca da solução do problema.

- O problema
  - Conjunto de dados:
    - Aspecto intensional: estrutura ou esquema do conjunto de dados;
    - Aspecto extensional: os casos ou registros.
  - Geralmente o processo KDD pressupõe uma estrutura matricial bidimensional – representando dados n-dimensionais.
  - Data Warehouse são ambientes úteis (mas não são requeridos necessariamente)

Caracterização

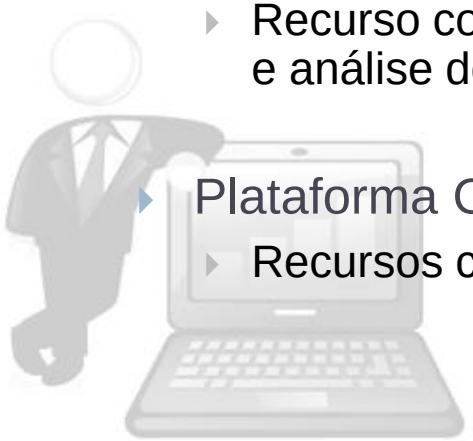
- O especialista no domínio da aplicação:
  - Pessoas ou grupos de pessoas
    - Analistas de negócios
    - Influenciam o processo em termos de estabelecimento de objetivos e avaliação de resultados.

- Os objetivos da aplicação:
  - Características esperadas do modelo de conhecimento a ser produzido ao final do processo
    - Restrições e expectativas
      - Exemplo: Precisão Mínima
  - Podem não estar bem definidos no início do processo e devem ser refinados.



Caracterização

- Os recursos disponíveis para a solução do problema:
  - O especialista em KDD:
    - Pessoa ou grupo de pessoas que possui experiência na execução de processos de KDD
    - Devem:
      - Identificar e utilizar conhecimento *a priori* sobre o problema
      - Escolher ferramentas e métodos
      - Direcionar as ações do processo
      - Conduzir a avaliação dos resultados
  - Ferramenta KDD
    - Recurso computacional que possa ser utilizado no processo de tratamento e análise de dados
  - Plataforma Computacional
    - Recursos computacionais de hardware

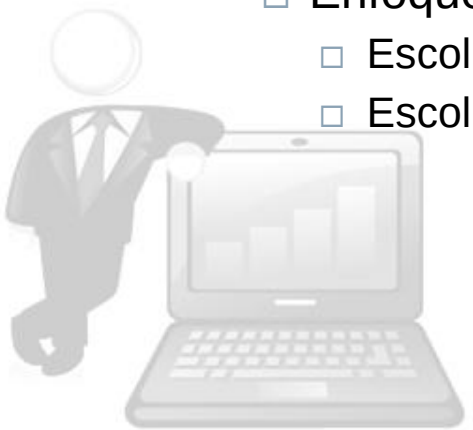


Caracterização

- Os resultados obtidos
  - Modelo de conhecimento
    - Indica qualquer abstração de conhecimento, expresso em alguma linguagem, que descreva algum conjunto de dados (Fayyad).
    - Vários modelos podem ser obtidos.
  - Histórico
    - Sobre como os modelos de conhecimento foram gerados.
    - Permitem reprodução, análise e revisão do processo.

Caracterização

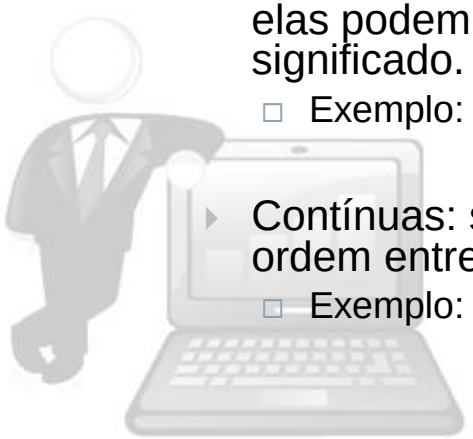
- Funções de Pré-Processamento:
  - Seleção de Dados
    - Identificação de quais dados das bases de dados existentes, devem ser efetivamente considerados durante o processo KDD
      - Exemplos:
        - O nome do cliente é irrelevante quando o objetivo e prever comportamento de novos clientes quanto ao pagamento de créditos;
        - A idade do cliente é fundamental quando o objetivo é estimar o valor de uma apólice de seguros.
      - Enfoques:
        - Escolha de atributos
        - Escolha de registros



Detalhando

Variáveis do problema (ou atributos)

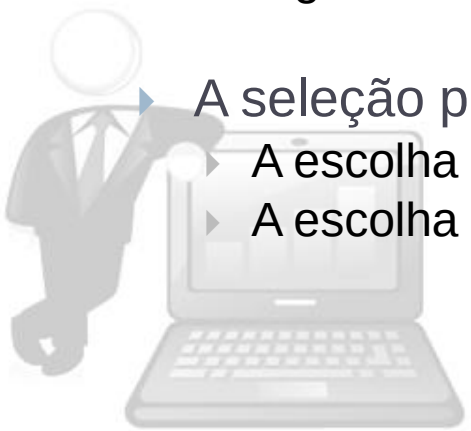
- Tipos de dados: indicam a forma em que eles estão armazenados.
- Tipos de variáveis: expressam a natureza com que a informação deve ser interpretada.
  - Classificação:
    - Nominais ou Categóricas: atribuem rótulos aos objetos. Podem assumir valores pertencentes a um conjunto finito e pequeno de estados possíveis. Não há ordenamento de seus valores.
      - Exemplo: estado civil
    - Discretas: Assemelham-se às variáveis nominais, mas os valores (estados) que elas podem assumir possuem um ordenamento, e este possui algum significado.
      - Exemplo: dias da semana
    - Contínuas: são variáveis quantitativas cujos valores possuem uma relação de ordem entre eles, podendo ser finito ou infinito.
      - Exemplo: renda



Tarefas de pré-processamento

Detalhando

- Seleção: quais informações devem ser efetivamente consideradas durante o processo de KDD.
  - Geralmente exige a realização da uma operação de junção de dados em uma única matriz bi-dimensional:
    - Junção direta: todos os atributos e registros da base de dados transacional são incluídos na nova estrutura (matriz), sem uma análise crítica quanto a que variáveis e que casos podem realmente contribuir para o processo de KDD.
    - Junção orientada: o especialista do domínio da aplicação, em pareceria com o especialista em KDD, escolhe os atributos e os registros com algum potencial para influenciar no processo de KDD.
  - A seleção pode ter dois enfoques distintos:
    - A escolha de registros
    - A escolha de atributos



Peres, S. M.

Redução de Dados Horizontal

Detalhando

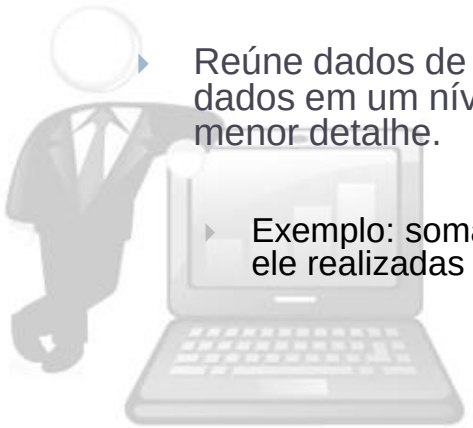
- Segmentação de Banco de Dados

```
SELECT *
FROM Cliente
WHERE tp_res = “P”;
```

- Eliminação Direta de Casos

```
DELETE
FROM Cliente
WHERE tp_res <> “P”;
```

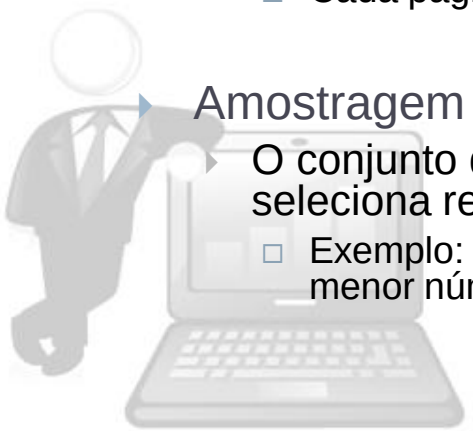
- Agregação de Informações
  - Reúne dados de forma a reduzir o conjunto de dados original. Na agregação de informações, dados em um nível maior de detalhamento são consolidados em novas informações com menor detalhe.
    - Exemplo: somar os valores de todas as compras de cada cliente, obtendo o total de despesas por ele realizadas durante um determinado período.



Redução de Dados Horizontal

Detalhando

- Amostragem Aleatória
  - Amostragem Aleatória Simples sem Reposição:
    - Todas as tuplas tem a mesma probabilidade de seleção e não existe reposição de uma tupla já selecionada.
  - Amostragem Aleatória Simples com Reposição:
    - Todas as tuplas tem a mesma probabilidade de seleção mas existe a reposição de uma tupla já selecionada.
  - Amostragem de Clusters (Agrupamento):
    - As tuplas devem ser agrupadas em M grupos (clusters) de tal forma que possa ser realizada uma amostragem aleatória entre os grupos.
      - Cada página do banco de dados pode ser considerada um grupo.
  - Amostragem Estratificada:
    - O conjunto de dados deve ser dividido em grupos disjuntos e a amostragem seleciona representantes de cada grupo.
      - Exemplo: clientes agrupados por faixa etária: mesmo os clientes da faixa etária com menor número de elementos serão representados na amostra final.



Redução de Dados Vertical

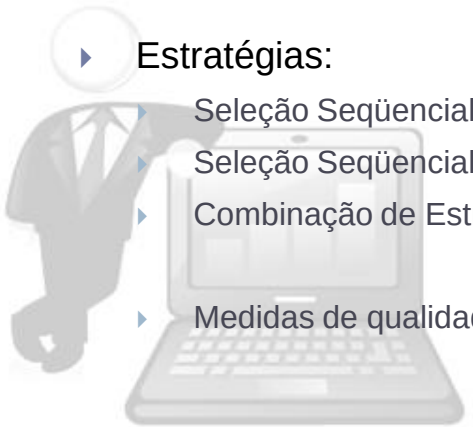
- Sendo *S* um conjunto de dados com atributos A_1, A_2, A_3, ..., A_n, o problema da redução de dados vertical consiste em identificar quais combinações entre tais atributos devem ser consideradas no processo de descoberta de conhecimento.

- Motivações:
  - Maior concisão e precisão;
  - Maior eficiência em termos de tempo de execução;
  - **É mais significativa em termos de redução do tamanho de um conjunto de dados do que a redução de dados horizontal.**

- Abordagens:
  - Independente do Modelo
  - Dependente do Modelo: experimenta o algoritmo de mineração para cada conjunto de atributos e avalia os resultados.

- Estratégias:
  - Seleção Seqüencial para a Frente (Forward Selection)
  - Seleção Seqüencial para Trás (Backward Selection)
  - Combinação de Estratégias

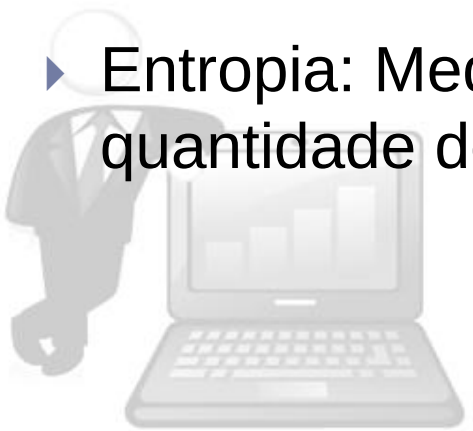
- Medidas de qualidade: entropia, dimensão fractal, taxa de inconsistências.



Redução de Dados Vertical

Detalhando

- Taxa de Inconsistências: é gerada a partir do agrupamento dos registros que possuem os mesmos valores com relação aos atributos em análise (exemplo pag. 31).
- Dimensão Fractal: análise da real porção do espaço ocupada por um objeto.
- Entropia: Medida de organização, de energia, de quantidade de informação.

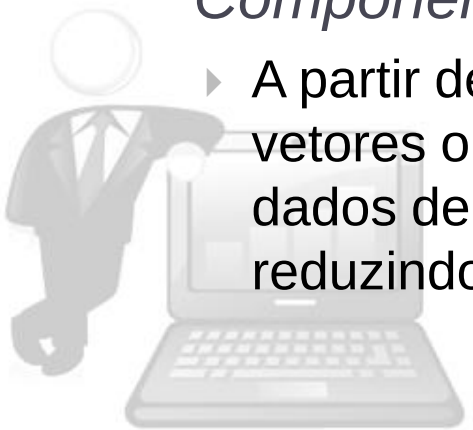


Peres, S. M.

Redução de Dados Vertical

Detalhando

- Métodos
  - Eliminação Direta de Atributos
    - Refere-se à eliminação de atributos cujo conteúdo não seja relevante ao processo de KDD.
      - Heurísticas
        - Eliminar todos os atributos que apresentem valores constantes em todos os registros.
        - Eliminar os atributos que sejam identificadores.
  - Análise de Componentes Principais (*Principal Componentes Analysis*)
    - A partir de dados normalizados a PCA encontra uma base de vetores ortonormais que permite fazer um mapeamento dos dados dentro das dimensões onde a variância é maior, reduzindo o número de atributos mas explicando bem os dados.



Redução de Valores

Detalhando

- Redução de Valores:
  - alternativa à opção de corte de atributos → reduz o número de valores distinto minimizando quantidade de comparações nos algoritmos.

  - Nominais
    - Identificação de Hierarquia entre Atributos
      - Logradouro, bairro, cidade e estado → podemos usar apenas o estado, por exemplo.
    - Identificação de Hierarquia entre Valores
      - Tênis, sandália, sapato → calçados
      - Bermuda, calça, camisa → roupas.

  - Contínuos (ou Discretos)
    - Particionamento em Células de mesma Cardinalidade
    - Redução de Valores pelas Medianas das Células
    - Redução de Valores pelas Médias das Células
    - Redução de Valores pelos Limites das Células
    - Arredondamento de Valores
    - Agrupamento de Valores

Redução de Valores

Detalhando

- Particionamento em Células de mesma Cardinalidade: os valores são agrupados em células e essas são substituída por identificadores.

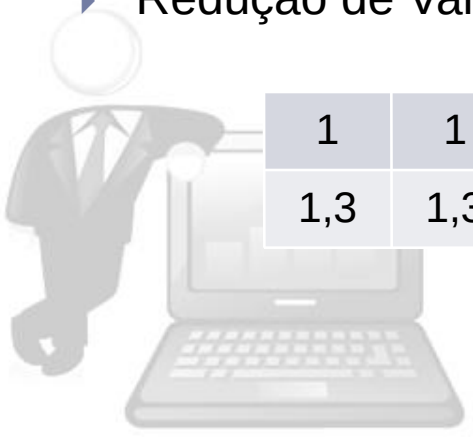
1	1	2	3	3	3	4	5	5	7
Id1	Id1	Id1	Id2	Id2	Id2	Id3	Id3	Id3	Id3

- Redução de Valores pelas Medianas das Células: idem o anterior mas com uso das medianas.

1	1	2	3	3	3	4	5	5	7
1	1	1	3	3	3	5	5	5	5

- Redução de Valores pelas Médias das Células: usando as médias ....

1	1	2	3	3	3	4	5	5	7
1,3	1,3	1,3	3	3	3	5,3	5,3	5,3	5,3



Redução de Valores

Detalhando

- Redução de Valores pelos Limites das Células ...

1	1	2	3	3	3	4	5	5	7
1	1	2	3	3	3	4	4	4	7

- Arredondamento de Valores:
  - utilização de algum procedimento de arredondamento ou aproximação de valores.
- Agrupamento de Valores:
  - agrupa valores de um atributo em clusters levando em consideração a similaridade existente entre tais valores.
    - depois de executado um processo de clusterização, cada clusters pode ser representado pela média dos valores a ele atribuídos.

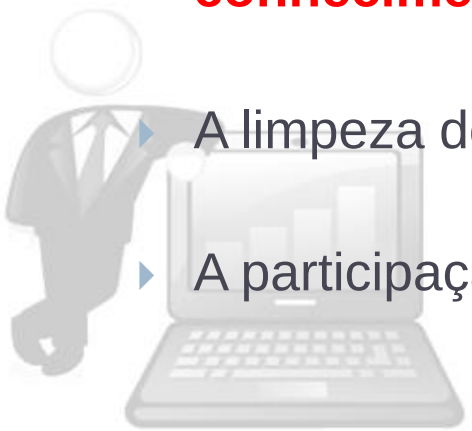
Caracterização - pré-processamento

- Limpeza de Dados
  - Tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos por eles representados.
  - Correções de informações ausentes, errôneas ou inconsistentes.
    - Exemplo:
      - Definição de um intervalo para um determinado atributo. Medidas de correção para registros com ocorrência fora do intervalo para o atributo.
      - Padronização de unidades.

Tarefas de pré-processamento

Detalhando

- Limpeza:
  - Em aplicações reais é comum encontrar dados:
    - Incompletos: se há informação ausente para determinados atributos ou ainda se há dados pouco detalhados.
    - Ruidosos: dados errados ou que contenham valores considerados divergentes (outliers) do padrão esperado.
    - Inconsistentes: contêm algum tipo de discrepância semântica entre si.
  - **Quanto pior for a qualidade dos dados informados ao processo KDD, pior será a qualidade dos modelos e conhecimento gerados.**
  - A limpeza dos dados objetiva melhorar a qualidade dos mesmos.
  - A participação do especialista do domínio, nesta fase, é essencial.

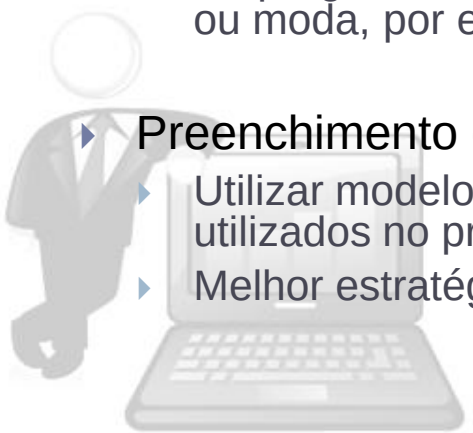


Peres, S. M.

Limpeza de Informações Ausentes

Detalhando

- Exclusão de Casos
  - Excluir do conjunto de dados as tuplas que possuam pelo menos um atributo não preenchido
- Preenchimento Manual de Valores
  - Preencher o que falta com base em pesquisas nas fontes originais dos dados
- Preenchimento com Valores Globais Constantes
  - Substituir todos os valores ausentes de um atributo por um valor padrão (“desconhecido” ou “null”), especificado pelo especialista de domínio.
- Preenchimento com Medidas Estatísticas
  - Empregar medidas estatísticas como alternativa à utilização de constantes (média ou moda, por exemplo).
- Preenchimento com Métodos de Mineração de Dados
  - Utilizar modelos preditivos para sugerir os valores mais prováveis a serem utilizados no preenchimento dos valores ausentes.
  - Melhor estratégia (as demais podem ser tendenciosas demais)



Limpeza de Inconsistências

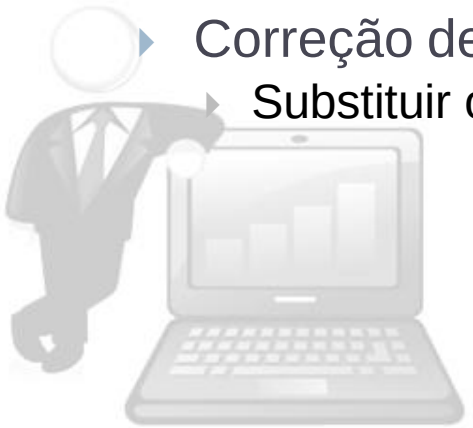
Detalhando

- Uma inconsistência pode envolver uma única tupla ou um conjunto de tuplas. Demanda conhecimento especialista.
  - Um cliente com idade inferior a 21 possui crédito aprovado.

- Métodos:

  - Exclusão de Casos
    - Excluir do conjunto de dados original as tuplas que possuem pelo menos uma inconsistência.
    - SQL pode ser utilizada para encontrar tais tuplas (regras de negócio).

  - Correção de Erros
    - Substituir os valores errôneos / corrigir as inconsistências.



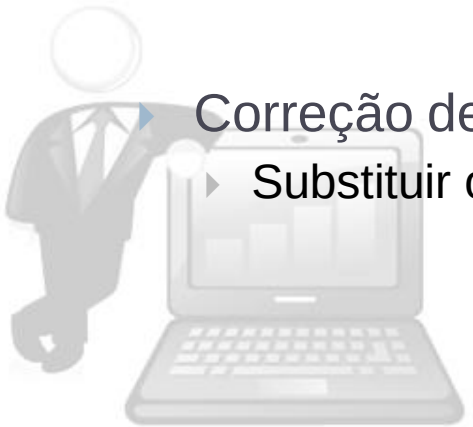
Limpeza de Valores não pertencentes ao Domínio

- Identificação e eliminação de valores que não pertencem ao domínio dos atributos do problema.
  - Caso particular da limpeza de inconsistências
  - Demanda conhecimento especialista
    - O valor “T” não pertence ao atributo “sexo”.

- Métodos

  - Exclusão de Casos
    - Excluir do conjunto de dados original, as tuplas que apresentam pelo menos um valor fora do conjunto de valores válidos.

  - Correção de Erros
    - Substituir os valores inválidos.



Caracterização - pré-processamento

- Codificação dos Dados:
  - Os dados devem ser codificados para ficarem numa forma que possam ser usados como entrada dos algoritmos de mineração:
    - Numérica-Categórica
    - Categórica-Numérica

Codificação

Detalhando

- Numérica-Categórica
  - Mapeamento Direto: substituição de valores numéricos por valores categóricos.
    - Sexo:
      - 1 → M
      - 0 → F
  - Mapeamento em Intervalos (Discretização): divisão do domínio de uma variável numérica em intervalos. (pode ser considerada como Redução de Valores)
    - Exemplo: considere o atributo renda com os seguintes valores, já organizados em ordem crescente → 1000, 1400, 1500, 1700, 2500, 3000, 3700, 4300, 4500, 5000.
    - Divisão em intervalos com comprimentos definidos pelo usuário:

Intervalo	Número de valores no intervalo
1000 |- 1600	3
1600 |- 4400	5
4400 |- 5400	2

Codificação

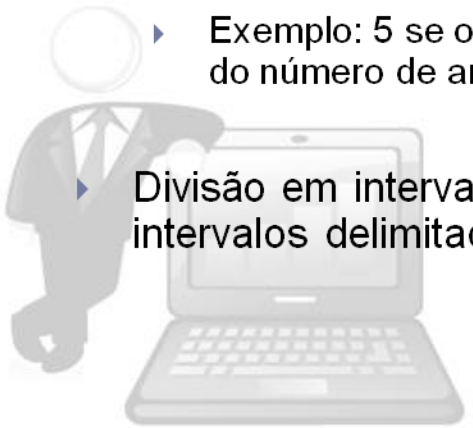
- Divisão em intervalos de igual comprimento: o usuário define apenas o número de intervalos e o comprimento dos intervalos é calculado a partir do maior e do menor valor do domínio do atributo.
  - 4 intervalos. Como a faixa de valores vai de 1000 a 5000, faz-se R = 5000 – 1000 = 4000. Assim, cada intervalo terá comprimento 1000 (4000/4)

Intervalo	Número de valores no intervalo
1000 |- 2000	4
2000 |- 3000	1
3000 |- 4000	2
4000 |- 5000	3

  - Critérios para definição do número de intervalos.
    - Exemplo: 5 se o número de amostras for < 25, senão o número de intervalos = raiz quadrada do número de amostras.

Heurística!!!

- Divisão em intervalos por meio de clusterização: cada clusters representa um intervalos delimitado pelo menor e maior valor no cluster.



Detalhando

Codificação

- Categórica-Numérica
  - Representação Binária Padrão (Econômica)

Valores Originais	Representação Binária Padrão
Casado	001
Solteiro	010
Viúvo	100
Divorciado	011
Outra	110

  - Representação Binária 1-de-N: o código tem um comprimento igual ao número de categorias discretas

Valores Originais	Representação Binária Padrão
Casado	00001
Solteiro	00010
Viúvo	00100
Divorciado	01000
Outra	10000

Codificação

- Representação Binária por Temperatura: quando os valores discretos estão relacionados de algum modo.
  - A diferença entre os conceitos “fraco” e “forte” deve ser a maior possível.
  - As diferenças entre valores adjacentes deve ser a menor possível;

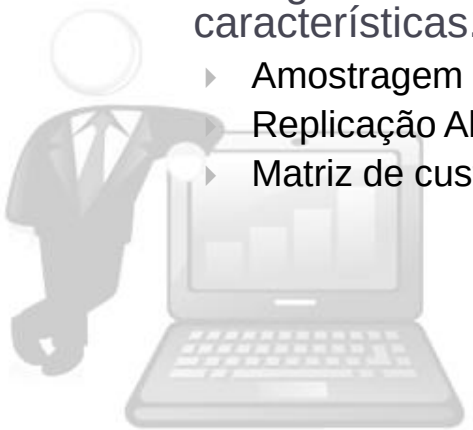
Valores Originais	Representação Binária Padrão
Fraco	0001
Regular	0011
Bom	0111
Forte	1111

Caracterização - pré-processamento

- Enriquecimento dos Dados:
  - Obter informações que possam ser agregadas em registros existentes.
  - Necessidade de pesquisas para complementação dos dados, consultas a bases externas ...
  - Exemplo:
    - Renda, Despesas, Tipo de Residência, Bairro de Residências
    - Renda, Despesas, Tipo de Residência, Bairro de Residências, Valor Médio de Imóvel

Caracterização - pré-processamento

- Normalização
  - Ajustar a escala de valores de cada atributo de forma que os valores fiquem em pequenos intervalos, tais como -1 e 1 ou 0 e 1.
    - Procedimentos pg. 45-49.
- Construção de atributos
  - Gerar novos atributos a partir de atributos existentes (derivação)
- Correção de Prevalência
  - Corrigir um eventual desequilíbrio na distribuição de registros com determinadas características.
    - Amostragem Estratificada
    - Replicação Aleatória de Registro
    - Matriz de custo: pesos associados aos erros ocorridos com classes menos numerosas.

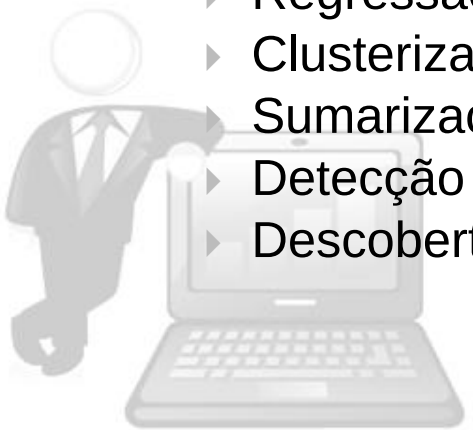


Caracterização - pré-processamento

- Partição do conjunto de dados
  - Conjunto de treino e conjunto de teste.
    - Holdout: esse método divide aleatoriamente os registros em uma percentagem fixa *p* para treinamento e (1 – *p*) para teste, considerando normalmente *p* > ½.
    - K-Fold CrossValidation: esse método consiste em dividir aleatoriamente o conjunto de dados com N elementos em K subconjuntos distintos, com aproximadamente o mesmo número de elementos. Neste processo, cada um dos K subconjuntos é utilizado como conjunto de teste e os (K – 1) demais subconjuntos são reunidos em um conjunto de treinamento. Vários modelos são gerados.
    - Stratifield K-Fold CrossValidation: similar a anterior, mas cada subconjunto deverá conter uma quantidade proporcional ao número de dados nas classes do conjunto de dados (para classificação).
    - Leave-One-Out: similar ao anterior, mas cada um dos K subconjuntos possui apenas um registro.
    - Bootstrap: o conjunto de treinamento é construído por meio de sorteios com reposição e os conjuntos de teste são feitos com os dados não sorteados para o conjunto de treinamento.

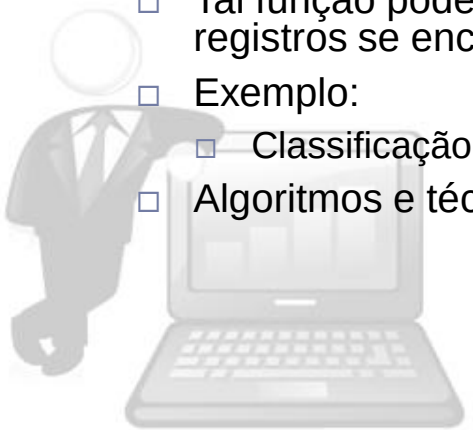
Caracterização

- Mineração (principal etapa);
  - Envolve a escolha e aplicação de uma técnica para efetiva produção de conhecimento.
  - Geralmente dependente do tipo de tarefa de KDD a ser realizada.
  - Tarefas:
    - Descoberta de Associação
    - Classificação
    - Regressão
    - Clusterização
    - Sumarização
    - Detecção de Desvios
    - Descoberta de Seqüências



Tarefas

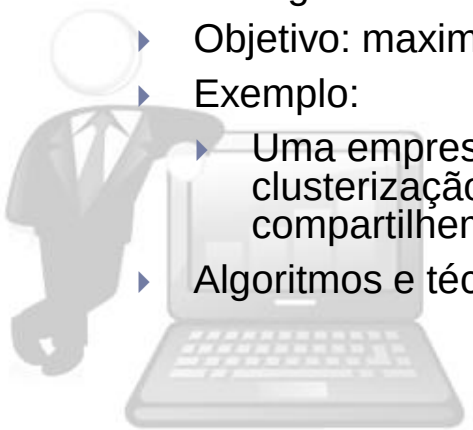
- Descoberta de Associação:
  - Abrange a busca por itens que freqüentemente ocorram de forma simultânea em transações do banco de dados.
  - Muito utilizada/solicitada pela área de marketing das empresas.
  - Exemplo clássico:
    - A associação descoberta entre fraldas e cervejas em uma rede de supermercados.
  - Algoritmos e técnicas: Apriori, GSP e DHP.
- Classificação:
  - Consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos predefinidos, denominados classes.
  - Tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se encaixam.
  - Exemplo:
    - Classificação de clientes em mal-pagadores e bom-pagadores.
  - Algoritmos e técnicas: Redes Neurais Artificiais, Algoritmos Genéticos, Lógica Indutiva ...



Tarefas

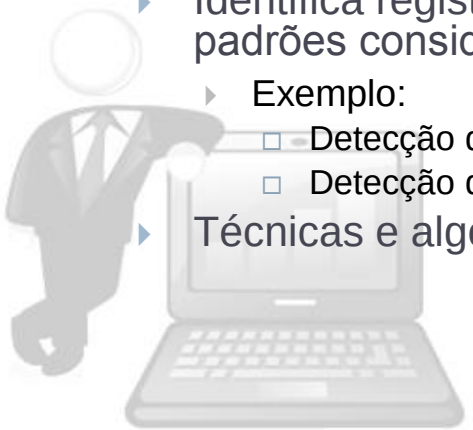
- Regressão
  - Busca por uma função que mapeie os registros de um banco de dados em valores reais.
  - Exemplo:
    - Predição da soma da biomassa presente em uma floresta;
    - Estimativa da probabilidade de um paciente sobreviver dado o resultado de um conjunto de diagnósticos de exames;
    - Predição do risco de determinados investimentos, de limite de cartão de crédito ....
  - Algoritmos e técnicas: Métodos Estatísticos, Redes Neurais Artificiais ...

- Clusterização (ou Agrupamento)
  - Separa os registros de uma base de dados em subconjuntos ou clusters (ou grupos), de tal forma que os elementos de um cluster compartilhem de propriedades comuns que os distingam de elementos de outros clusters.
  - Objetivo: maximizar a similaridade intraclusters e minimizar a similaridade interclusters.
  - Exemplo:
    - Uma empresa do ramo de telecomunicações pode realizar um processo de clusterização de sua base de clientes de forma a obter grupos de clientes que compartilhem o mesmo perfil de compra.
  - Algoritmos e técnicas: K-means, K-Modes, Redes Neurais Artificiais ...



Tarefas

- Sumarização:
  - Consiste em procurar identificar e indicar características comuns entre conjuntos de dados.
  - Exemplo:
    - Considere um banco de dados com informações sobre clientes que assinam um determinado tipo de revista semanal. A sumarização deve buscar por características que seja comuns a boa parte dos clientes.
  - É comum aplicar sumarização a cada um dos agrupamentos obtidos na tarefa de clusterização
  - Algoritmos e técnicas: Lógica Indutiva, Algoritmos Genéticos.
- Detecção de Desvios:
  - Identifica registros de um banco de dados cujas características não atendam aos padrões considerados normais no contexto. Tais registros são os "*outliers*".
    - Exemplo:
      - Detecção de comportamento estranho de clientes de uma operadora de cartão de crédito
      - Detecção de intrusão em uma rede de computadores
  - Técnicas e algoritmos: Estatística, Sistemas Imunológicos Artificiais ...



Tarefas

- Descoberta de Seqüências:
  - Extensão da tarefa de descoberta de associações em que são buscados itens freqüentes considerando várias transações ocorridas ao longo de um período.
    - Exemplo
      - Considerando o exemplo das compras no supermercado. Se o banco de dados possui a identificação do cliente associada a cada compra, a tarefa de descoberta de associação pode ser ampliada de forma a considerar a ordem em que os produtos são comprados ao longo de tempo.
  - Algoritmos e técnicas: Estatística, Redes Neurais Artificiais ...

Caracterização

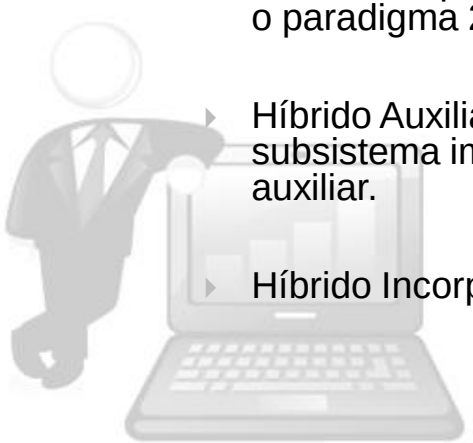
- Pós-Processamento:
  - Tratamento do conhecimento obtido
  - Nem sempre necessária
  - Tem o objetivo de facilitar a interpretação e a avaliação, pelo homem, da utilidade do conhecimento descoberto.
  - Envolve:
    - A organização da informação através de:
      - Gráficos
      - Diagramas
      - Relatórios
    - A conversão da forma de representação do conhecimento obtido.

Caracterização

- Classificando KDD:
  - Quanto à orientação das ações a serem realizadas:
    - Validação de hipótese
    - Descoberta de conhecimento
  - Quanto ao macroobjetivo desejado:
    - Predição
    - Descrição

Caracterização

- Técnicas Tradicionais:
  - Existem independentemente do contexto de Mineração de Dados.
- Técnicas Específicas:
  - Desenvolvidos especificamente para aplicação em tarefas de KDD.
- Técnicas Híbridas:
  - Combinação de técnicas para resolução de um problema
    - Híbrido Seqüencial: um sistema com paradigma 1 atua como entrada de outro sistema com o paradigma 2.
    - Híbrido Auxiliar: um subsistema constituído pela técnica do paradigma 2 é chamado pelo subsistema implementado pelo paradigma 1, retornando ou realizando alguma tarefa auxiliar.
    - Híbrido Incorporado: não há uma separação nítida entre dois subsistemas.



Caracterização

- O ser humano como elemento central do processo de KDD.