CIÊNCIA DE DADOS - 01

Prof. Júlio Cesar Nievola

PPGIa – PUCPR

06/abril/2019

Termos relacionados

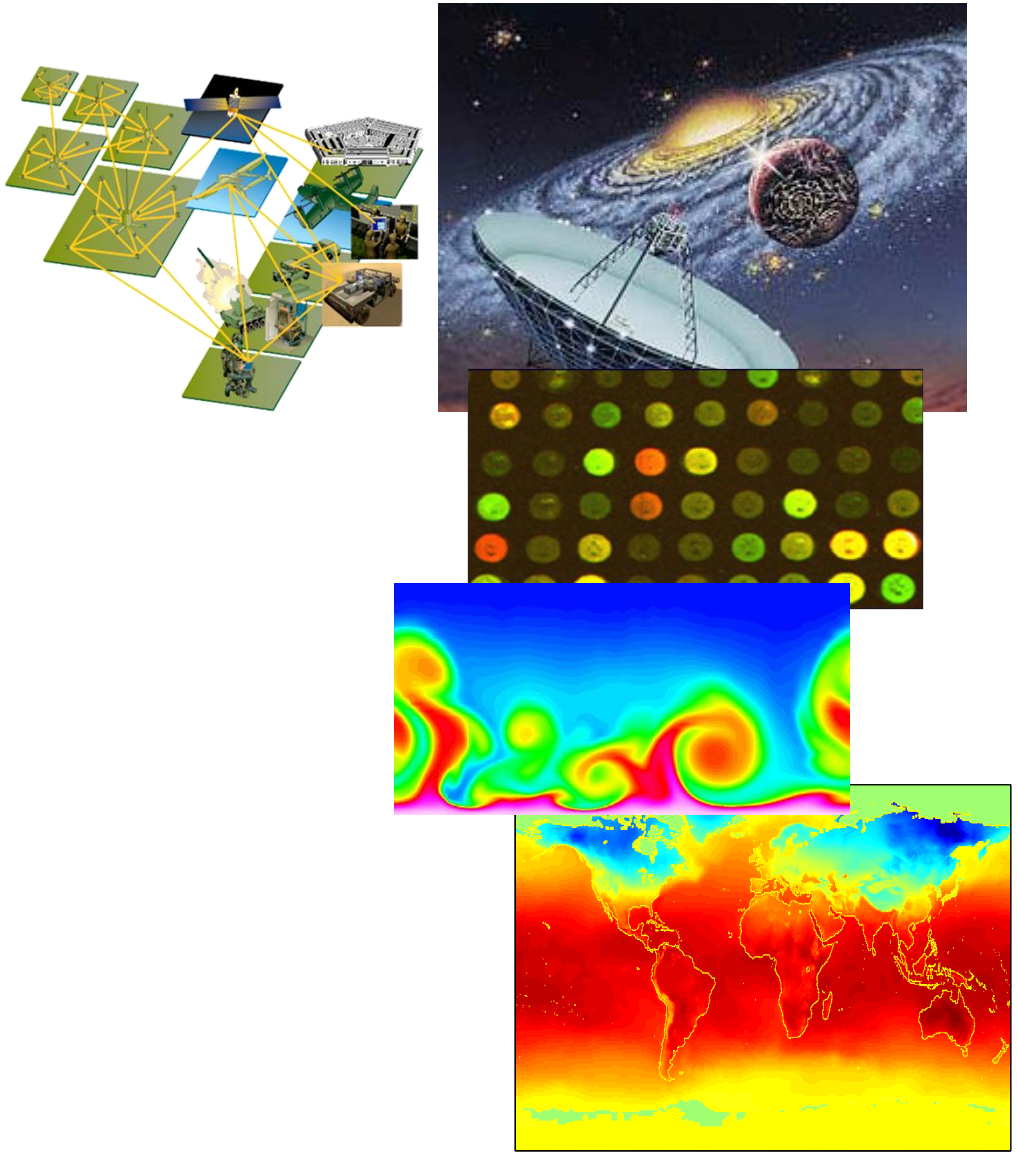
Por quê se preocupar com dados? Ponto de vista comercial

- Muitos dados são coletados e armazenados
  - Web data, e-commerce
  - Compras em departamentos/ supermercados
  - Bancos / Transações com cartão de crédito

- Computadores se tornaram baratos e [illegible]
- Pressão competitiva é forte
  - Fornecer serviços melhores e personalizados como um diferencial (e.g. em CRM)

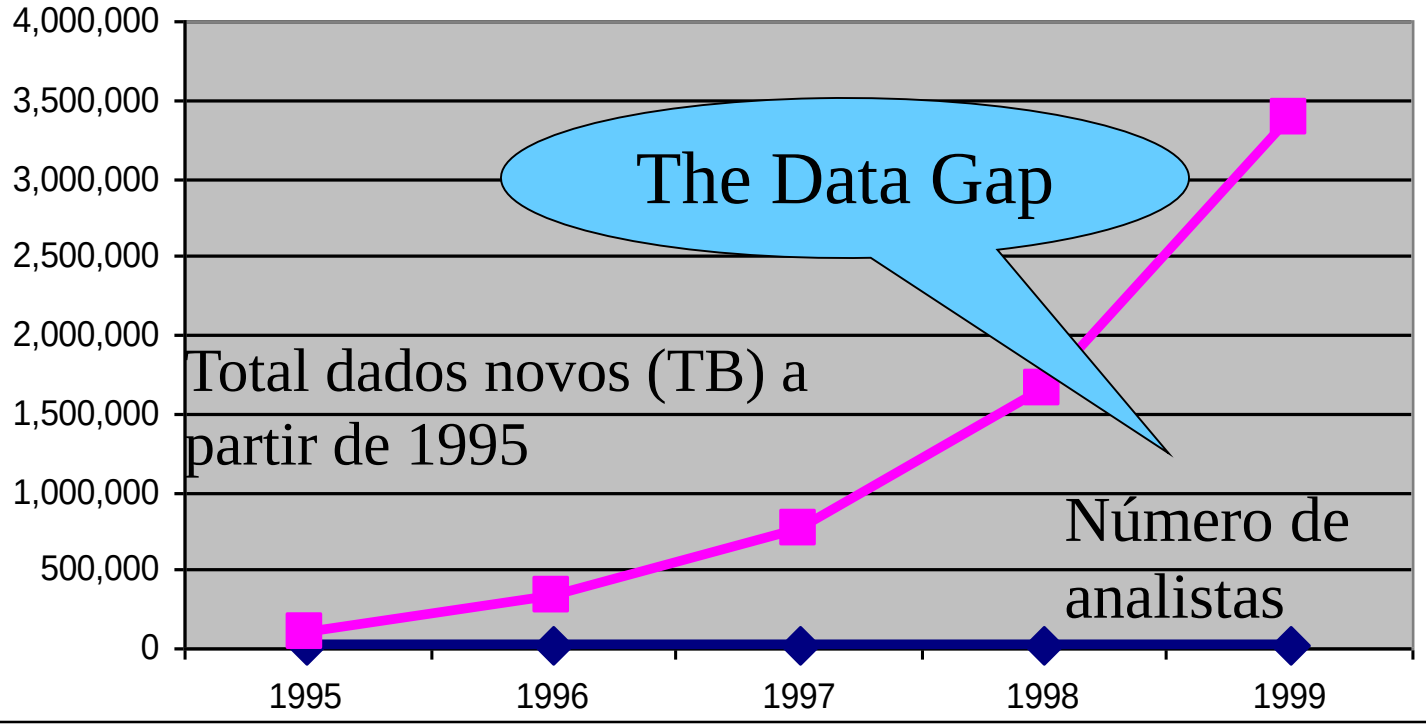
Por quê se preocupar com dados? Ponto de vista científico

- Dados coletados a enormes velocidades (GB/hora)
  - Sensores remotos em satélites
  - Telescópios sondando o céu
  - Micro-arranjos gerando dados de expressão gênica
  - Simulações científicas gerando terabytes de dados
- Técnicas tradicionais inviáveis para dados brutos
- Mineração de dados pode ajudar cientistas
  - classificando e segmentando dados
  - na Formulação de Hipóteses



Trabalhando com Grandes Bases de Dados

- Frequentemente há informação "escondida" nos dados que não está prontamente evidente
- Analistas humanos podem levar semanas para descobrir informação útil
- Muito dos dados não é analisada nunca



Referência: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Cadeia de Valores

Dados

- Dados de clientes
- Dados armazenados
- Dados demográficos
- Dados geográficos

Informação

- X vive em Z
- S tem Y anos
- X e S mudaram de local
- W tem dinheiro em Z

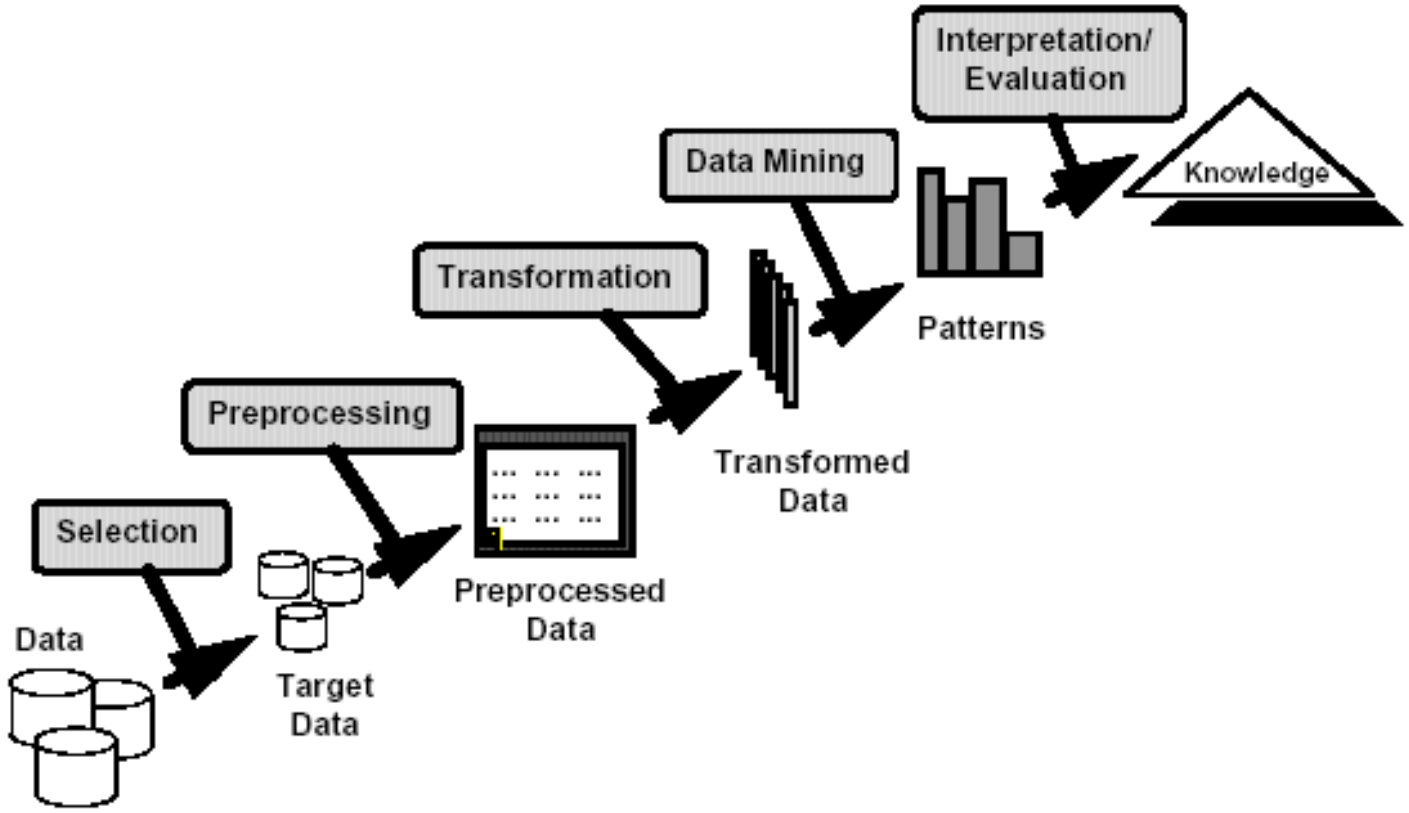
Conhecimento

- Uma quantidade Y do produto A é usado na região Z
- Clientes da classe Y usam x% de C no período D

Decisão

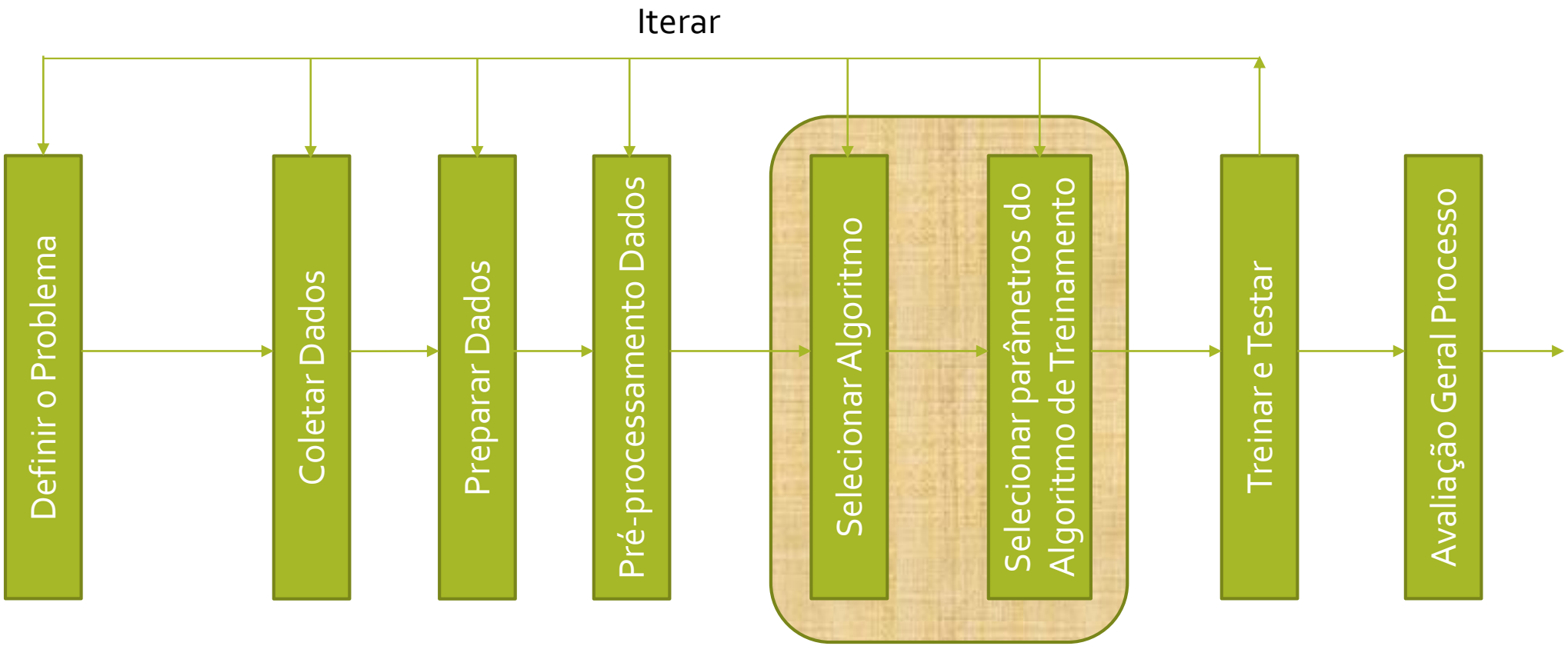
- Promover produto A na região Z.
- Enviar propaganda para famílias de perfil P
- Vender serviço B para clientes C

Processo KDD



Referência: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., Advances in Knowledge Discovery and Data Mining, 1996

Processo em Ciência de Dados



Tipos de Tarefas a partir dos Dados

- Tarefa Preditiva
  - Usa algumas variáveis para prever valores desconhecidos ou futuros de outras variáveis

- Tarefa Descritiva
  - Encontra padrões compreensíveis por humanos para descrever os dados

Detalhamento dos Tipos de Tarefas

- Classificação [Preditiva]
- Agrupamento [Descritiva]
- Descoberta de Regras de Associação [Descritiva]
- Descoberta de Padrões Seqüenciais [Descritiva]
- Regressão [Preditiva]
- Detecção de Desvios [Preditiva]

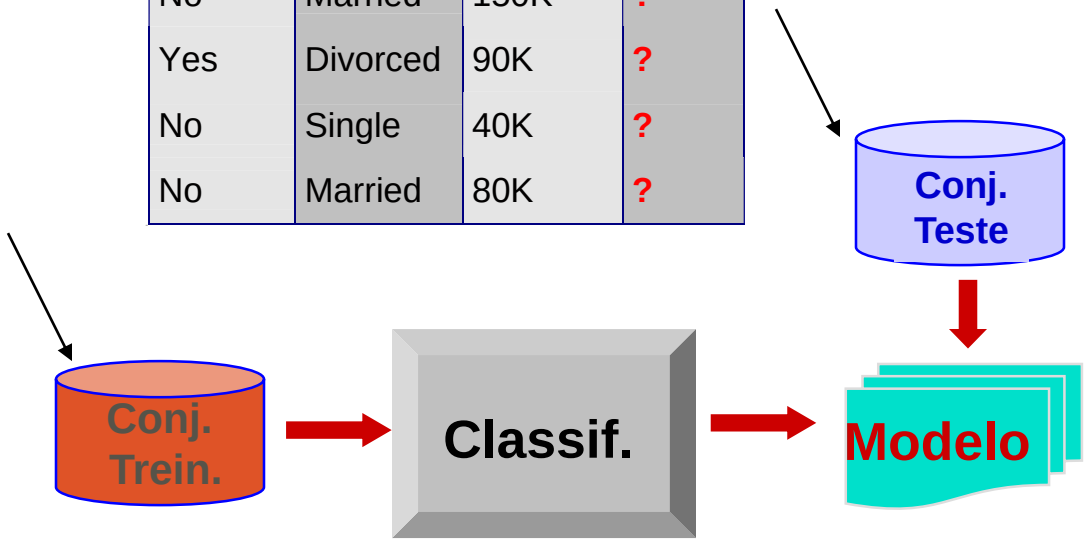
Classificação: Definição

- Dado um conjunto de registros (*conjunto de treinamento*)
  - Cada registro contém um conjunto de *atributos*, um dos atributos é a *classe*.
- Encontrar um *modelo* para o atributo classe como uma função dos valores dos outros atributos.
- Objetivo: a registros previamente não-usados deve ser assinalada uma classe tão precisamente quanto possível.
  - Um *conjunto de testes* é usado para determinar a precisão do modelo. Usualmente, o conjunto de dados é dividido em conjunto de treinamento e conjunto de testes, sendo o conjunto de treinamento usado para construir o modelo e o conjunto de testes usado para validá-lo.

Classificação: Exemplo

	categórico	categórico	contínuo	classe
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**



Classificação: Aplicação 1

- Marketing Direto
  - Objetivo: Reduzir custo de propaganda *escolhendo* um conjunto de clientes que provavelmente comprarão um novo produto celular.
  - Abordagem:
    - Usar os dados de maneira similar ao exemplo anterior.
    - Sabe-se quais clientes decidiram comprar o produto e quais não. Esta decisão *{comprar, não-comprar}* forma o *atributo classe*.
    - Coletar várias informações demográficas, de estilo de vida, e de interação com a empresa relacionadas a todos os clientes.
      - Tipo de negócio, onde eles ficam, quanto recebem, etc.
    - Usar esta informação como atributos de entrada para treinar um modelo de um classificador.

Classificação: Aplicação 2

- Detecção de Fraude
  - Objetivo: Prever casos fraudulentos em transações de cartão de crédito.
  - Abordagem:
  - Usar transações de cartão de crédito e a informação sobre os clientes como atributos.
    - Quando um cliente compra, o que ele compra, quão frequentemente ele paga em dia, etc.
  - Rotular as transações passadas como transação do tipo fraude ou honesta. Isto forma o atributo classe.
  - Treinar um modelo para a classe das transações.
  - Usar este modelo para detectar fraude observando transações de cartão de crédito sobre uma conta.

Classificação: Aplicação 3

- Insatisfação de clientes:
  - Objetivo: Prever se um cliente tem propensão a migrar para um competidor.
  - Abordagem:
  - Usar registros detalhados de transações de cada um dos clientes passados e atuais, para encontrar atributos.
    - Quão frequentemente o cliente liga, para que setor ele liga, em que horário do dia ele liga mais, seu estado financeiro, estado civil, etc.
  - Rotular o cliente como leal ou não-leal.
  - Encontrar um modelo para a lealdade.

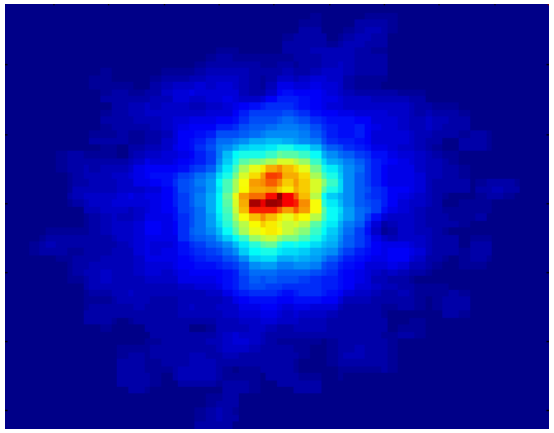
Classificação: Aplicação 4

- Catálogo de Pesquisa do Firmamento
  - Objetivo: Prever a classe (estrela ou galáxia) de objetos celestes, especialmente os visualmente muito fracos, baseado em imagens de pesquisa de telescópios (do Observatório Palomar).
    - 3000 imagens com 23,040 x 23,040 pixels por imagem.
  - Abordagem:
  - Segmentar a imagem.
  - Medir atributos da imagem (características) - 40 delas por objeto.
  - Modelar a classe baseado nestas características.
  - História de Sucesso: Encontrou 16 novos high red-shift quasars, alguns dos objetos mais distantes que são difíceis de encontrar!

Classificando Galáxias

Cortesia: http://aps.umn.edu

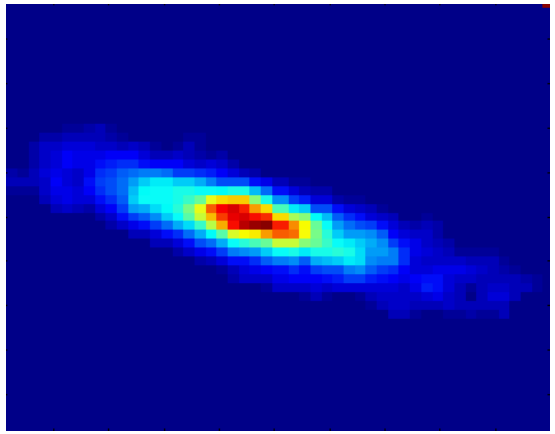
Inicial



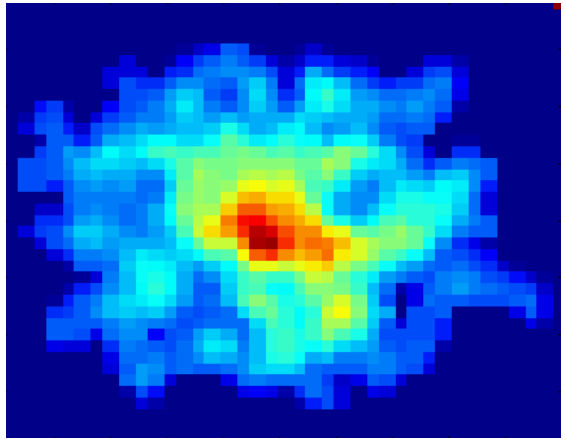
Classe:
- **Estágio de Formação**

Atributos:
- **Caract. da imagem,**
- **Características das ondas de luz recebidas, etc.**

Intermediário



Final



Quantidade de dados:
- **72 milhões de estrelas, 20 milhões de galáxias**
- **Catálogo de objetos: 9 GB**
- **Base de Dados das Imagens: 150 GB**

Tarefa de Agrupamento (“Clustering”)

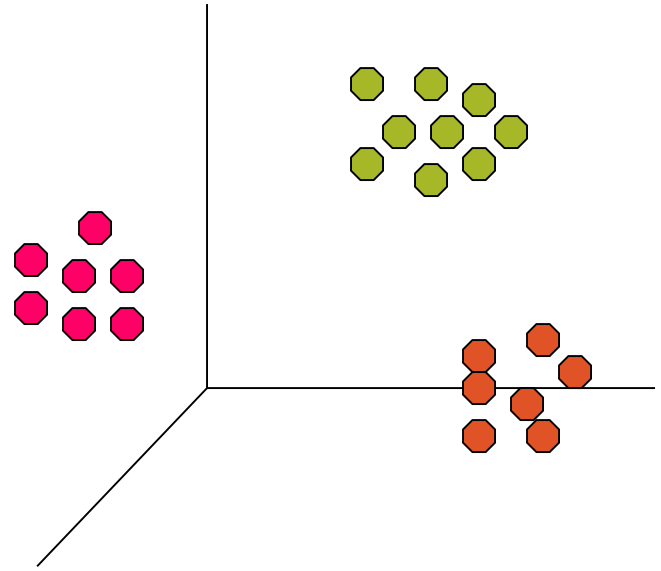
- Dado um conjunto de pontos de dados, cada um tendo um conjunto de atributos, e uma medida de similaridade entre eles, encontrar agrupamentos tais que
  - Pontos de dados em um grupo são mais similares entre si.
  - Pontos de dados em grupos diferentes são menos similares entre si.
- Medidas de Similaridade:
  - Distância Euclidiana se os atributos são contínuos.
  - Outras medidas dependentes do problema.

Ilustrando Agrupamento

- Agrupamento baseado em distância Euclidiana no espaço 3-D.

Distâncias intra-grupos são minimizadas

Distâncias inter-grupos são maximizadas



Agrupamento: Aplicação 1

- Segmentação de Mercado:
  - Objetivo: subdividir um mercado em distintos subconjuntos de clientes em que cada subconjunto pode ser visto como um mercado-alvo a ser atingido com uma mistura de marketing distintos.
  - Abordagem:
    - Coletar diferentes atributos de clientes baseado em informação relacionada ao seu estilo e posição geográfica.
    - Encontrar grupos de clientes similares.
    - Medir a qualidade dos grupos observando padrões de compra dos clientes no mesmo grupo versus aqueles de diferentes grupos.

Agrupamento: Aplicação 2

- Agrupamento de Documentos:
  - Objetivo: Encontrar grupos de documentos que são similares entre si baseado nos termos importantes que aparecem neles.
  - Abordagem: Identificar termos que ocorrem com frequência em cada documento. Formar uma medida de similaridade baseada na frequência dos diferentes termos. Usá-la para agrupar.
  - Ganho: Recuperação de Informações pode utilizar os grupos para relacionar um novo documento ou termo de pesquisa aos documentos agrupados.

Ilustrando Agrupamento de Documentos

- Pontos de Agrupamento: 3204 Artigos do Los Angeles Times.
- Medida de Similaridade: Quantas palavras são comuns nestes documentos (após alguma filtragem das palavras).

Categoria	*Total de Artigos*	*Corretamente colocados*
Financeiro	555	364
Estrangeiro	341	260
Nacional	273	36
Metrô	943	746
Esportes	738	573
Lazer	354	278

Agrupamento dos dados S&P 500

- Observar Stock Movements diários.
- Pontos de agrupamento: Stock-{UP/DOWN}
- Medida de Similaridade: Dois pontos são mais similares se os eventos que descrevem aparecem freqüentemente juntos nomesmo dia.
  - Usou-se regras de associação para quantificar a medida de similaridade.

	Grupos Descobertos	***Grupo Industrial***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Descoberta de Regras de Associação

- Dado um conjunto de registros, cada um dos quais contém certo número de itens de uma coleção;
  - Produzir regras de dependência que predirão a ocorrência de um item baseado nas ocorrências de outros itens.

ID	*Items*
1	Pão, Refri, Leite
2	Cerveja, Pão
3	Cerveja, Refri, Fralda, Leite
4	Cerveja, Pão, Fralda, Leite
5	Refri, Fralda, Leite

Regras Descobertas:

{Leite} --> {Refri}

{Fralda, Leite} --> {Cerveja}

Descoberta de Regras de Associação: Aplicação 1

- Marketing e Promoção de Vendas:
  - As regras descobertas tem o formato

    {Pão francês, … } --> {Batata frita}
  - Batata frita como conseqüente => Pode ser usada para determinar o que deve ser feito para incrementar as vendas.
  - Pão francês no antecedente => Pode ser usado para determinar quais produtos seriam afetados se a empresa descontinuasse a venda de pão frances.
  - Pão francês no antecedente *E* Batata frita no conseqüente => Pode ser usado para determinar quais produtos devem ser vendidos com Pão francês para promover a venda de Batatas fritas!

Descoberta de Regras de Associação: Aplicação 2

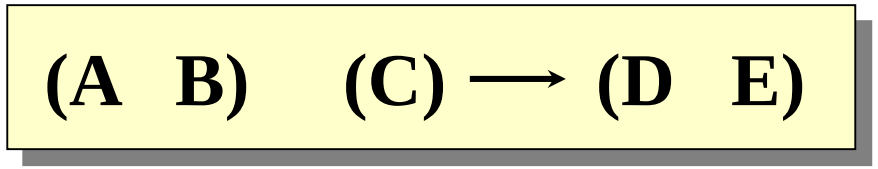
- Gerenciamento de prateleira de Supermercado.
  - Objetivo: Identificar itens que são comprados juntos por um número suficiente de clientes.
  - Abordagem: Processar os dados coletados no ponto-de-venda com scanners de código de barras para encontrar dependências entre itens.
  - Um regra clássica --
    - Se um cliente compra fraldas e leite, então ele provavelmente comprará cerveja.
    - Portanto, não se surpreenda se você encontrar engradados de cerveja próximos às fraldas!

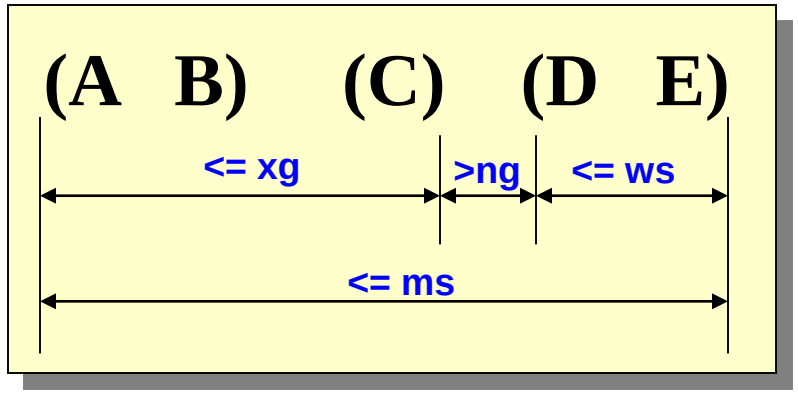
Descoberta de Regras de Associação: Aplicação 3

- Gerenciamento de Inventário:
  - Objetivo: Uma companhia de reparos de aparelhos domésticos quer antecipar a natureza dos reparos nos produtos de seus clientes e manter os veículos de serviço equipados com as partes certas para reduzir o número de visitas às casas dos clientes.
  - Abordagem: Processar os dados sobre ferramentas e partes necessárias em reparos prévios em diferentes localizações de clientes e descobrir os padrões de co-ocorrência.

Descoberta de Padrões Sequenciais: Definição

- Dado um conjunto de *objetos*, cada objeto associado com sua própria *linha do tempo de eventos*, encontrar regras que predigam fortes dependências sequenciais entre diferentes eventos.
- Regras são formadas descobrindo inicialmente padrões. As ocorrências de eventos nos padrões são governadas pelas restrições temporais.





Descoberta de Padrões Sequenciais: Exemplos

- Em logs de alarmes em telecomunicações,
  - (Inverter_Problem Excessive_Line_Current)

    (Rectifier_Alarm) ==> (Fire_Alarm)

- Em sequências de transações em pontos de venda,
  - Livrarias:

    (Intro_To_Visual_C) (C++_Primer) ==>
    (Perl_for_dummies,Tcl_Tk)
  - Loja de equipamentos de atletismo:

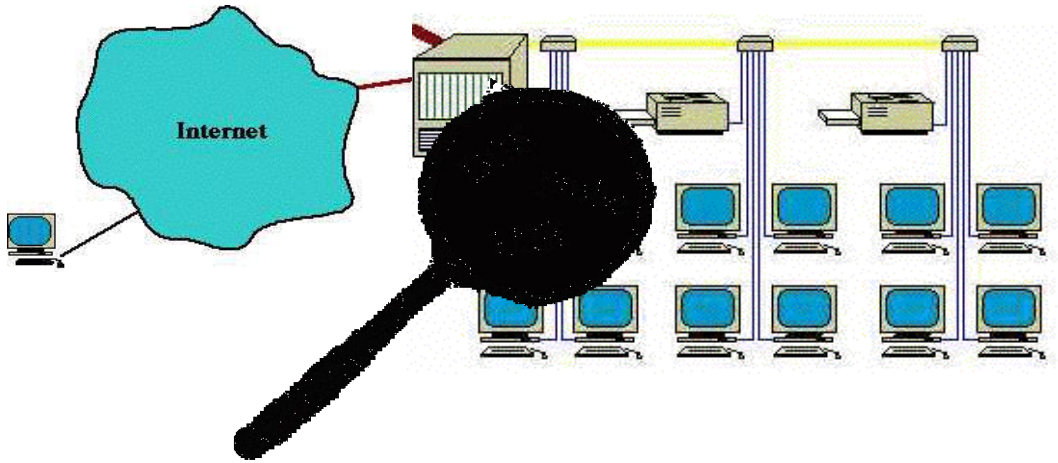
    (Shoes) (Racket, Racketball) ==> (Sports_Jacket)

Tarefa de Regressão

- Prevê um valor de uma dada variável continuamente valorada baseada nos valores de outras variáveis, assumindo um modelo de dependência linear ou não-linear.
- Muito estudado nos campos da estatística e redes neurais.
- Exemplos:
  - Prever quantidade de vendas de um novo produto baseado nos gastos de propaganda.
  - Prever velocidade dos ventos como uma função da temperatura, umidade, pressão do ar, etc.
  - Previsão de séries temporais de índices do mercado financeiro.

Detecção de Desvios / Anomalias

- Detectar desvios significantes do comportamento normal
- Aplicações:
  - Detecção de Fraudes em Cartões de Créditos
  - Detecção de Intrusão em Redes



Desafios em Mineração de Dados

- Escalabilidade
- Dimensionalidade
- Dados Complexos e Heterogêneos
- Qualidade dos Dados
- Propriedade e Distribuição dos Dados
- Preservação da Privacidade
- Fluxo de Dados

Bibliografia Básica

- Cohen, P., "Empirical Methods for Artificial Intelligence", The MIT Press, 1995.
- Pyle, D., "Data Preparation for Data Mining", Morgan Kaufmann Publishers, Inc., 1999.
- Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., "Data Mining – Practical Machine Learning Tools and Techniques", Morgan Kaufmann Publishers, Inc., 4th Edition, 2017.
- Janert, P.K., "Data Analysis with Open Source Tools", O'Reilly Media, Inc., 2011.
- Milton, M., "Head First Data Analysis", O'Reilly Media, Inc., 2009.