Mineração de Dados: Introdução

ACH5504 – Mineração de Dados

Notas de aulas baseadas no livro

"*Introduction to Data Mining*"

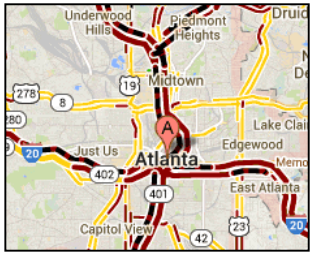
Tan, Steinbach, Karpatne, Kumar

Os dados em grande escala estão em todo lugar!

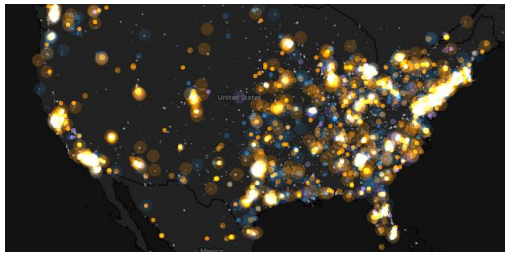
- Temos um enorme crescimento de quantidade de dados tanto em bases comerciais como científicas devido aos avanços na geração de dados e tecnologias de recolha
- A mantra nova
  - Recolhe os dados que puder quando e onde for possível.
- Expectativas
  - Os dados recolhidos terão valor para a finalidade recolhida ou para um propósito não previsto.

Segurança cibernética

E-Comércio

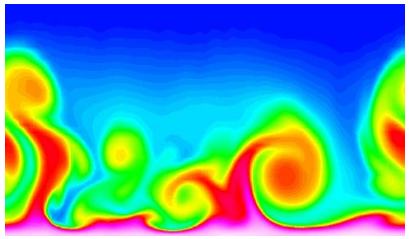


Padrões de transito



Rede Social: Twitter

Redes de sensores



Simulações computacionais

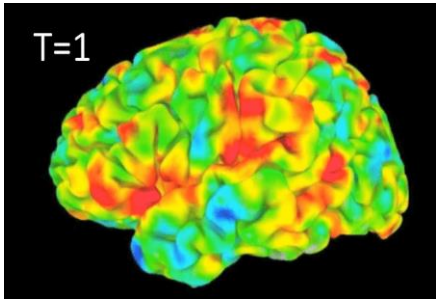
Porque a Mineração de Dados? Ponto de vista comercial

- Uma quantidade enorme de dados é coletada e armazenada
  - Dados da Web
    - Yahoo possui petabytes de dados web
    - Facebook possui biliões de usuários ativos
  - Compras nas lojas e supermercados, e-comércio
    - Amazon processa milhões de visitas por dia
  - transações bancarias e de cartões de crédito
- Os computadores tornaram-se mais baratos e mais poderosos para analisar os dados
- A pressão de competição é forte
  - Fornecer serviços melhores e mais customizados (e.g. na gestão de relacionamento com o cliente)

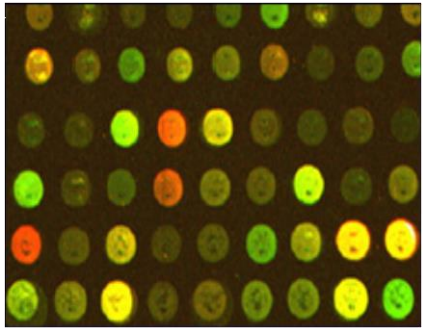
Por que a mineração de dados? Ponto de vista científico

- Dados recolhidos e armazenados com velocidades enormes
  - sensores remotos em um satélite
    - NASA EOSDIS arquiva mais do que um petabyte de dados da ciência da Terra/ano
  - telescópios escaneando o Céu
    - Sky survey data
  - Dados biológicos de alta taxa de transferência
  - simulações científicas
    - cerca de terabyte de dados gerados em poucas horas

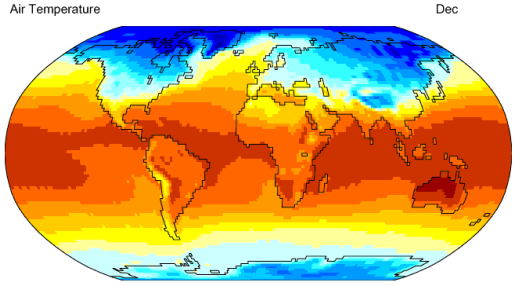
Sky Survey Data



Dados MRI do cérebro



Dados da expressão gênica



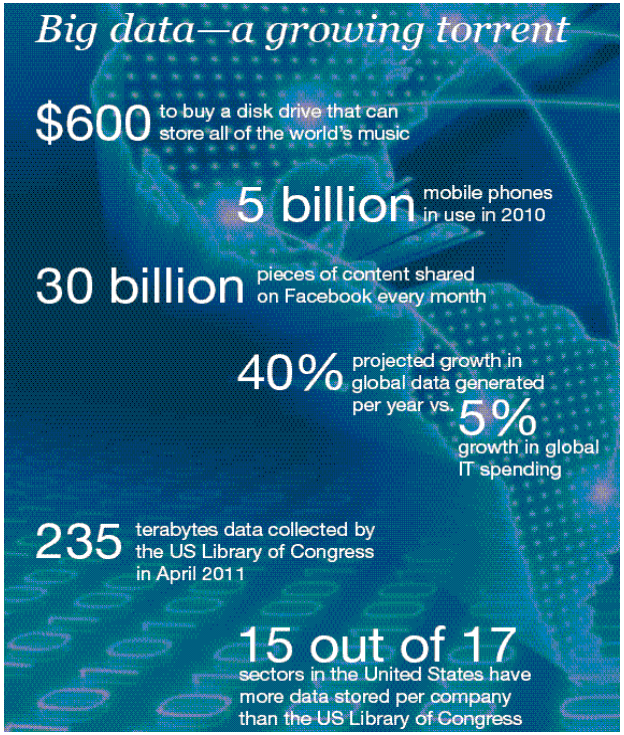
Temperatura da superfície da Terra

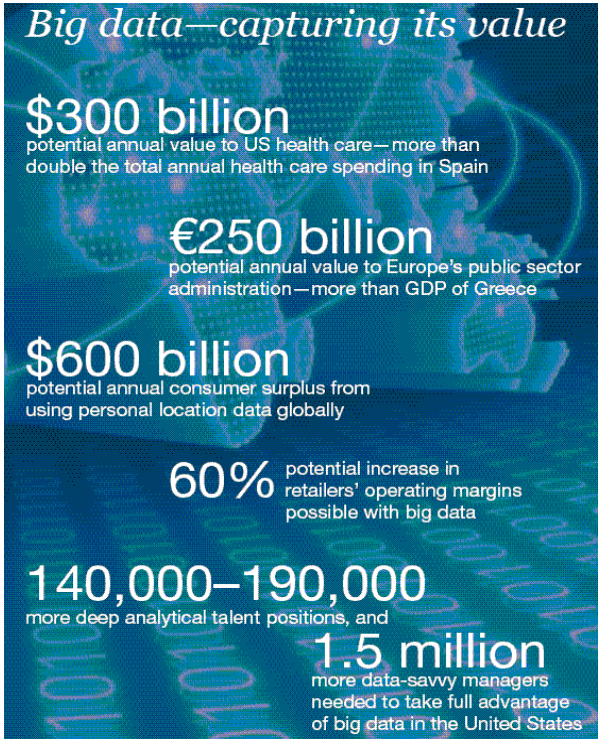
- A mineração de dados ajuda os cientistas
  - na análise automatizada de conjuntos de enorme quantidade de dados
  - na formação de hipóteses

Grandes oportunidades para melhorar a produtividade em todas as esferas da vida

McKinsey Global Institute

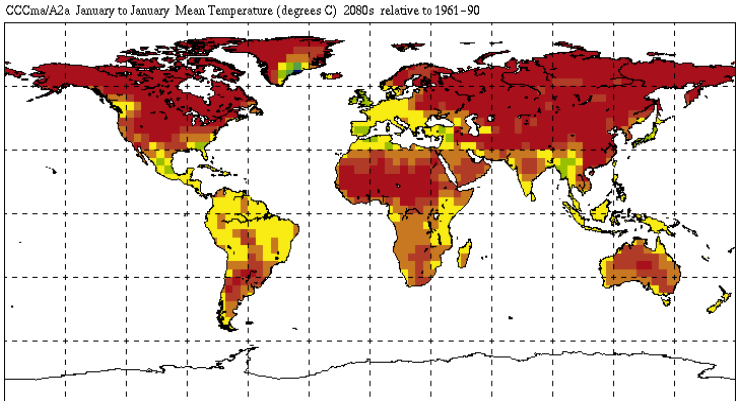
Big data: The next frontier for innovation, competition, and productivity





Grandes oportunidades para resolver os principais problemas da sociedade

Melhorar saúde e reduzir custos



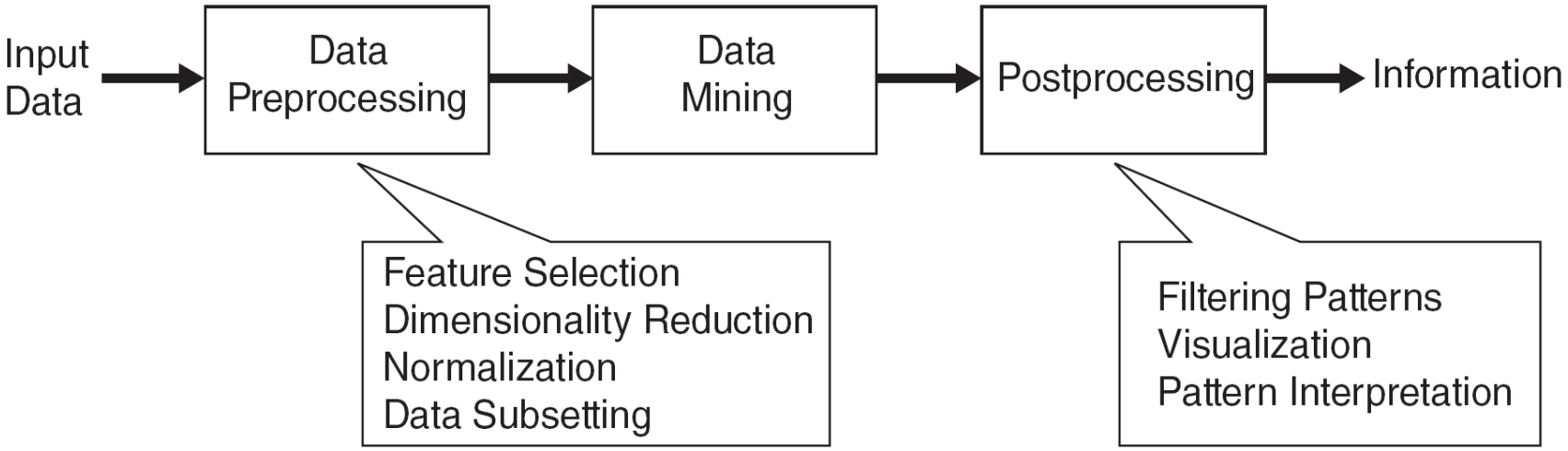
Prever o impacto das alterações climáticas

Encontrar fontes de energia alternativas/verdes

Aumentar a produção agrícola para reduzir a fome e a pobreza

O que é a Mineração de Dados?

- Várias definições
  - Extração não trivial de informações implícitas de dados, anteriormente desconhecidas e potencialmente úteis
  - Exploração e análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados, com objetivo de descobrir padrões significativos



O que (não) é a Mineração de Dados?

- **O que não é…**
  - Procurar o número de telephone nos contatos
  - Buscar a palavra “Amazon” no Google

- **O que é…**
  - Certos nomes são mais prevalentes em certos locais dos EUA (O’Brien, O’Rourke, O’Reilly… na área de Boston)
  - Agrupar documentos semelhantes devolvidos pelo motor de busca de acordo com o seu contexto (e.g., a floresta Amazônia, a empresa Amazon.com)

Historia da Mineração de Dados

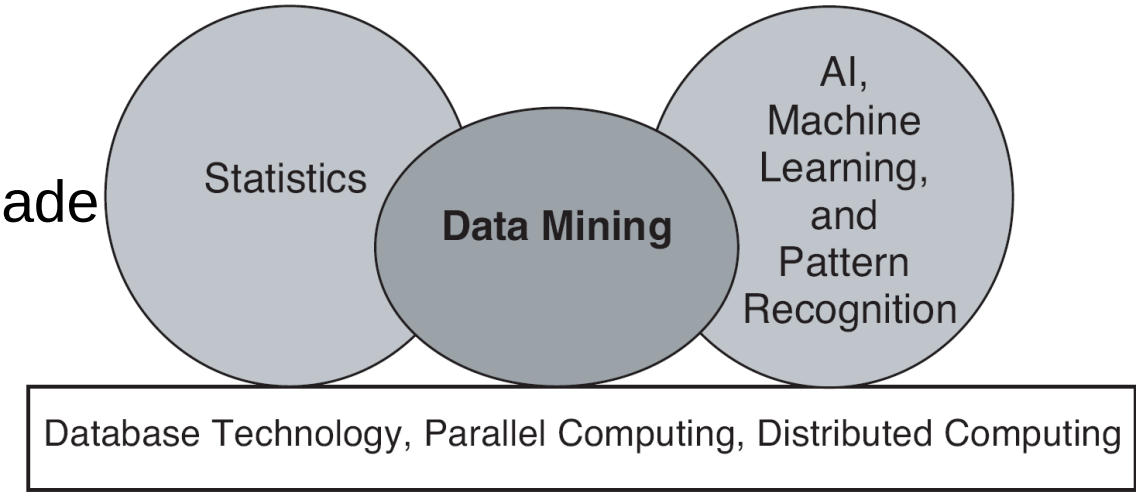
- **1763** Thomas Bayes publicou um artigo sobre um teorema para relacionar a probabilidade atual à probabilidade prévia chamada Teorema de Bayes. É fundamental para a mineração de dados e probabilidade, uma vez que permite a compreensão de realidades complexas com base em probabilidades estimadas.
- **1805** Adrien-Marie Legendre e Carl Friedrich Gauss aplicam regressão para determinar as órbitas de corpos sobre o Sol (cometas e planetas). O objetivo da análise de regressão era estimar as relações entre as variáveis, e o método específico que eles usaram neste caso era o método de mínimos quadrados. A regressão é uma das principais ferramentas de mineração de dados.
- **1936** Este é o inicio da idade de computador que faz possível a coleção e o processamento de grandes quantidades de dados. Em um artigo de 1936, Alan Turing introduziu a ideia de uma máquina universal capaz de realizar computações como nossos computadores modernos. O computador moderno é construído sobre os conceitos pioneiros do Turing.

Historia da Mineração de Dados

- **1943** Warren McCulloch e Walter Pitts foram os primeiros a criar um modelo conceitual de uma rede neural. Em um artigo intitulado “Um cálculo lógico das ideias imanentes na atividade nervosa”, eles descrevem a ideia de um neurônio em uma rede. Cada um desses neurônios pode fazer 3 coisas: receber entradas, processar entradas e gerar saída.
- **1975** John Henry Holland escreveu “Adaptação em sistemas naturais e artificiais”, o livro inovador sobre algoritmos genéticos. É o livro que iniciou este campo de estudo, apresentando as bases teóricas e explorando aplicações.
- **1989** O termo “Knowledge Discovery in Databases” (KDD) é inventado por Gregory Piatetsky-Shapiro. Também neste momento que ele organiza o primeiro workshop também chamado KDD.
- **2001** Embora o termo “Data Science” tenha existido desde 1960, o William S. Cleveland introduziu-o em 2001 como uma disciplina independente.

Origens da Mineração de Dados

- Empresta ideias de aprendizado de máquina/AI, reconhecimento de padrões, estatísticas e sistemas de banco de dados
- As técnicas tradicionais podem ser inadequadas devido a dados
  - de grande escala
  - de alta dimensionalidade
  - heterógenos
  - complexos
  - distribuidos



- Um componente chave do campo emergente da ciência de dados e da descoberta orientada por dados

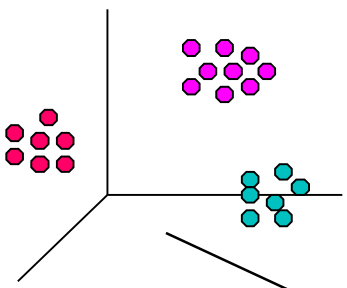
Tarefas de mineração de dados

- Métodos Preditivos
  - Use algumas variáveis para prever valores desconhecidos ou futuros de outras variáveis.

- Métodos Descritivos
  - Encontre padrões interpretáveis por humanos que descrevem os dados.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Tarefas de mineração de dados ...

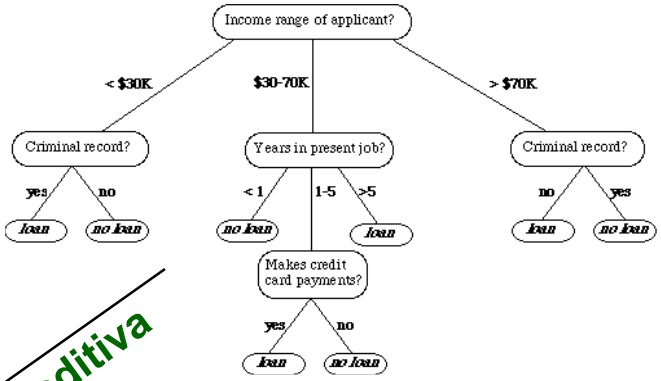


Clustering

Dados

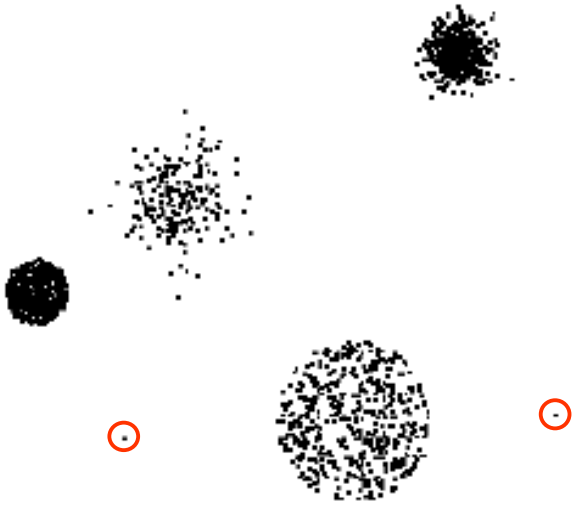
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Modelagem Preditiva



Regras de associação

Deteção de anomalias



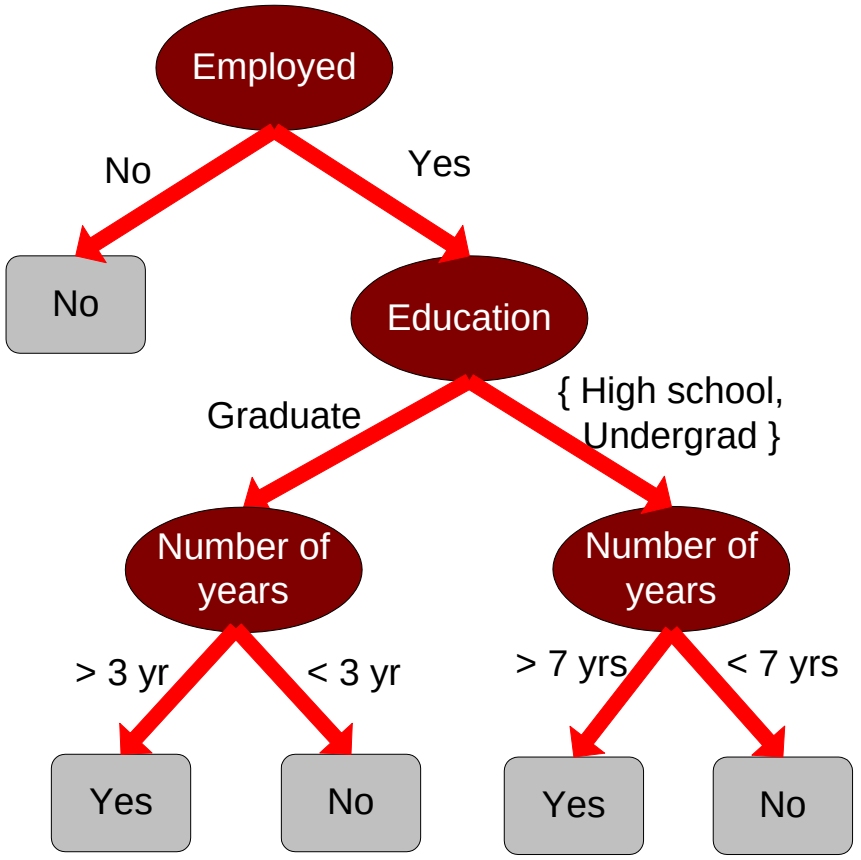
Modelagem preditiva: classificação

- Localizar um modelo para o atributo de classe como uma função dos valores de outros atributos

Modelo de aplicação de crédito/imprestimo

Classe

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

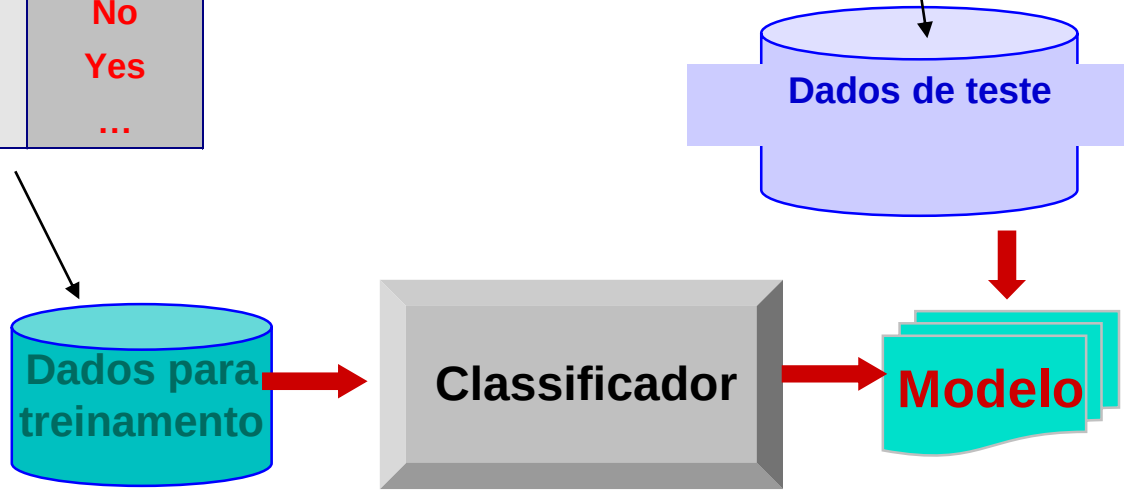


Exemplo de classificação

categórica | categórica | quantitativa | classe

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

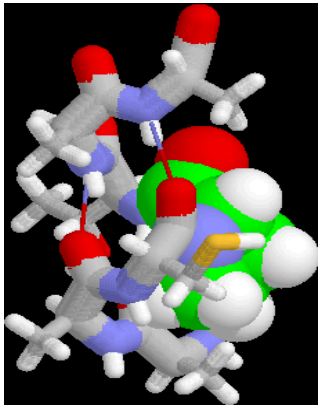


Exemplos de tarefas de classificação

- Classificação de transações de cartão de crédito como legítimas ou fraudulentas

- Classificação de coberturas terrestres (corpos hídricos, áreas urbanas, florestas, etc.) utilizando dados de satélites

- Categorizando notícias como finanças, tempo, entretenimento, esportes, etc.
- Identificando intrusos no ciberespaço
- Classificação de células tumorais como benignas ou malignas



- Classificação de estruturas secundárias das proteínas como a alfa-hélice, a beta-folha, ou a bobina aleatória

Classificação: aplicação 1

- Deteção de fraudes
- **Objetivo:** Prever casos fraudulentos em transações com cartão de crédito.
  - **Abordagem:**
    - Use transações de cartão de crédito e as informações do titular de conta como atributos.
      - Quando um cliente compra, o que ele compra, quantas vezes ele paga sem atraso, etc.
    - Marca transações passadas como legítimas ou fraude.
      - Isso forma o atributo de classe.
    - Aprenda um modelo para a classe das transações.
    - Use este modelo para detectar fraudes observando transações de cartão de crédito em uma conta.

Classificação: aplicação 2

- Previsão para clientes de telefonia que cancelarão o serviço
  - **Objetivo:** Prever se um cliente é susceptível de ser perdido para um concorrente.
  - **Abordagem:**
    - Use o registro detalhado de transações com cada um dos clientes passados e presentes, para encontrar atributos.
      - Quantas vezes o cliente chama, onde ele chama, que hora do dia que ele chama mais, seu status financeiro, estado civil, etc.
    - Rotule os clientes como leais ou desleais.
    - Encontre um modelo de fidelidade.

From [Berry & Linoff] Data Mining Techniques, 1997

Classificação: aplicação 3

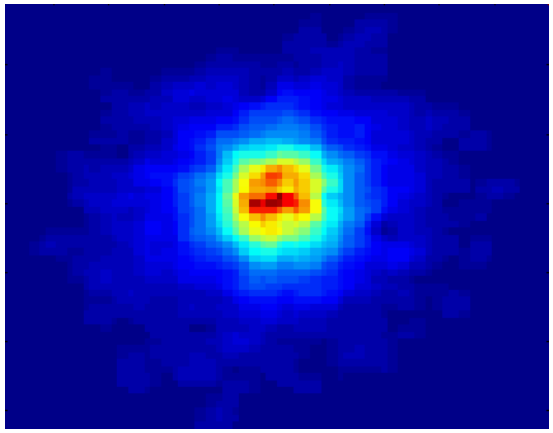
- Classificação de objetos de Sky Survey
  - **Objetivo:** Prever a classe (estrela ou galáxia) de objetos do céu, especialmente os visualmente fracos, com base nas imagens já classificadas (de Observatório de Palomar Observatory).
    - 3000 imagens com 23.040 x 23.040 píxeis por imagem.
  - **Abordagem:**
    - Segmentar a imagem.
    - Medir atributos de imagem (propriedades) - 40 por objeto.
    - Modelar a classe com base nesses atributos.
    - História de sucesso: poderia encontrar 16 novos quasars de alta red-shift, alguns dos mais distantes objetos que são difíceis de encontrar!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classificando as galáxias

Fonte: http://aps.umn.edu

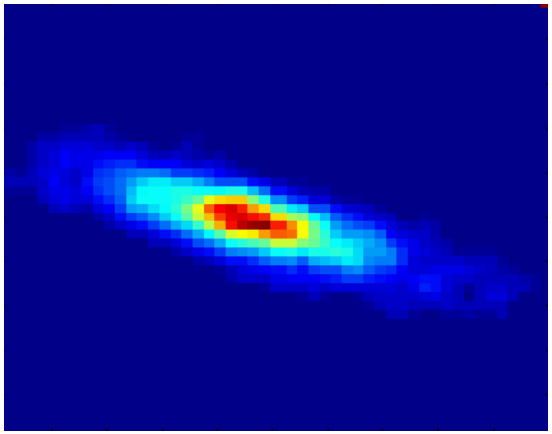
Cedo



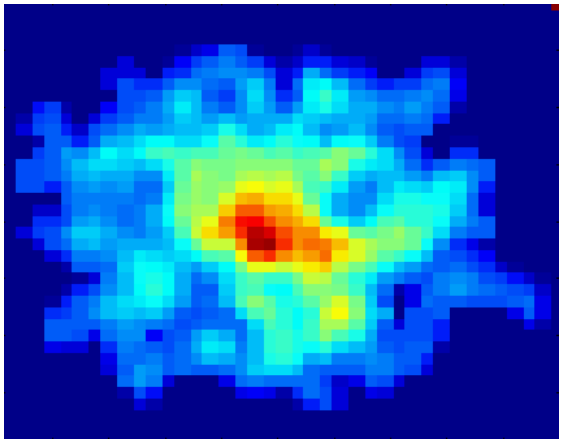
Class:
- **Estágios de formação**

Atributos:
- **Propriedades de imagen,**
- **Características da luz recebidas, etc.**

Intermediário



Tarde



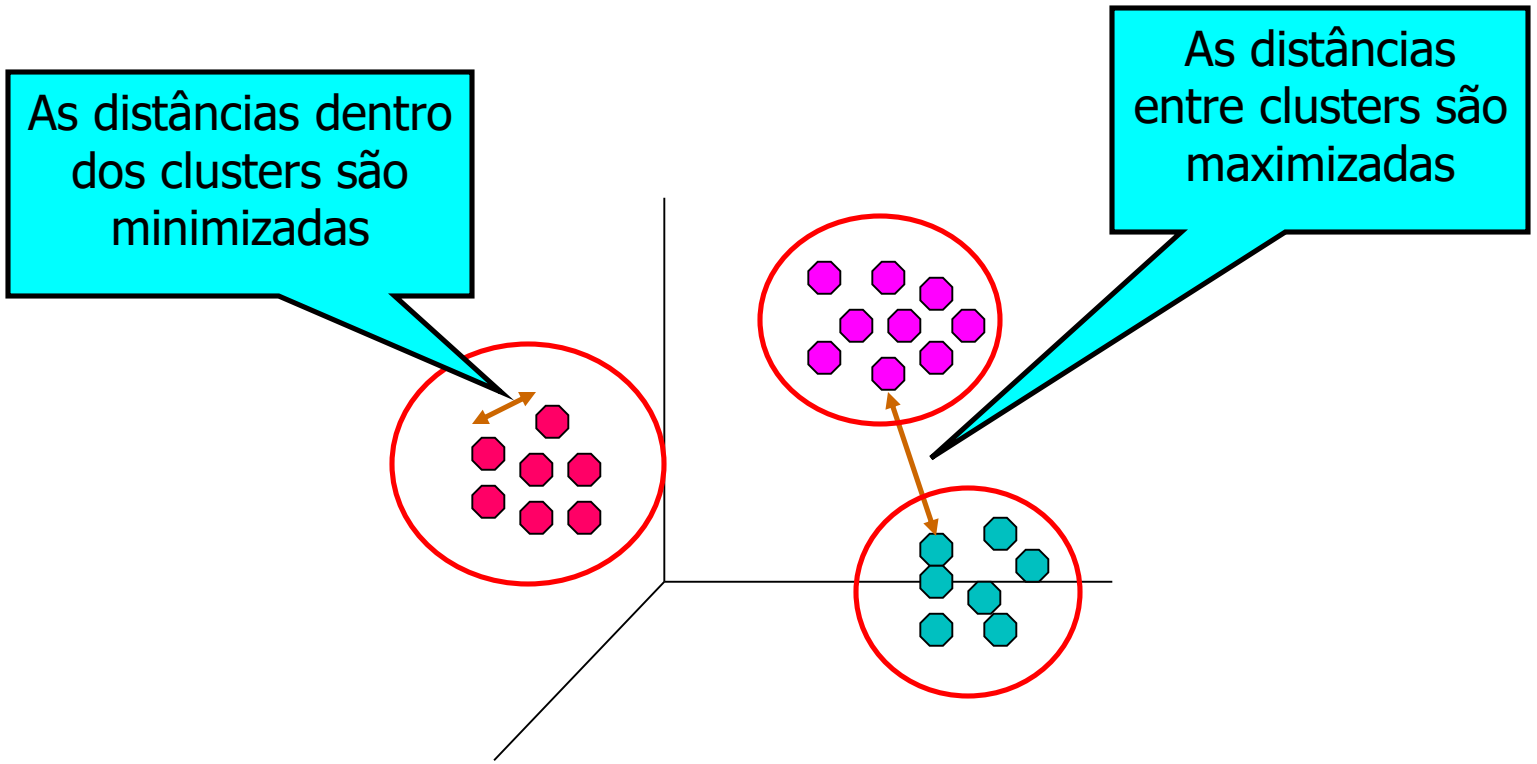
Tamanho dos dados:
- **72 milhões estrelas, 20 milhões galáxias**
- **Catálogo de objetos: 9 GB**
- **Banco de imagens: 150 GB**

Regressão

- Prever um valor de uma determinada variável com valor contínuo com base nos valores de outras variáveis, assumindo um modelo linear ou não linear de dependência.
- Extensivamente estudado em estatísticas, campos de redes neurais.
- Exemplos:
  - Prever quantidades de vendas de novos produtos com base em despesas de propaganda.
  - Previsão de velocidades de vento como uma função de temperatura, umidade, pressão de ar, etc.
  - Previsão de aumento/queda de índices nos mercado de ações.

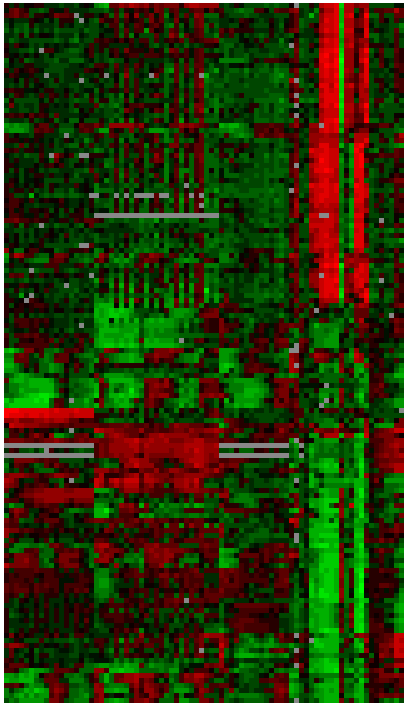
Clustering

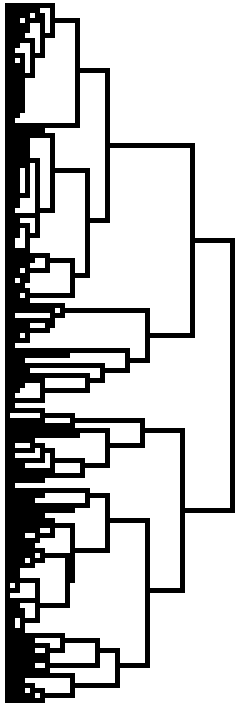
- Localizar grupos de objetos de tal forma que os objetos em um grupo serão semelhantes (ou relacionados) entre si e diferentes de (ou não relacionados a) os objetos em outros grupos



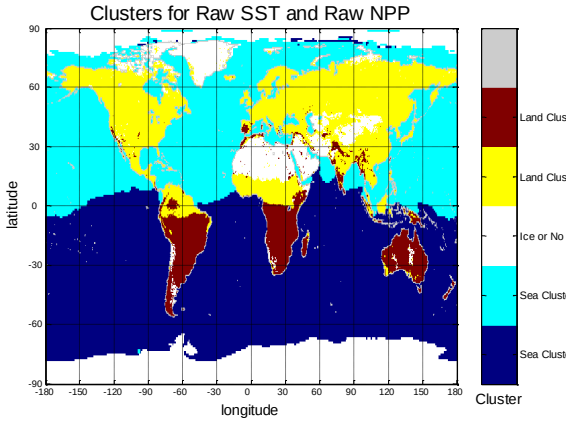
Aplicações de análise de clusters

- **Compreender**
  - Perfil do cliente para mercados
  - Agrupar documentos relacionados para navegação
  - Agrupar genes e proteínas que têm funcionalidade semelhante
  - Agrupar ações de mercado com variações de preço semelhantes
- **Resumir**
  - Reduzir o tamanho de grandes conjuntos de dados

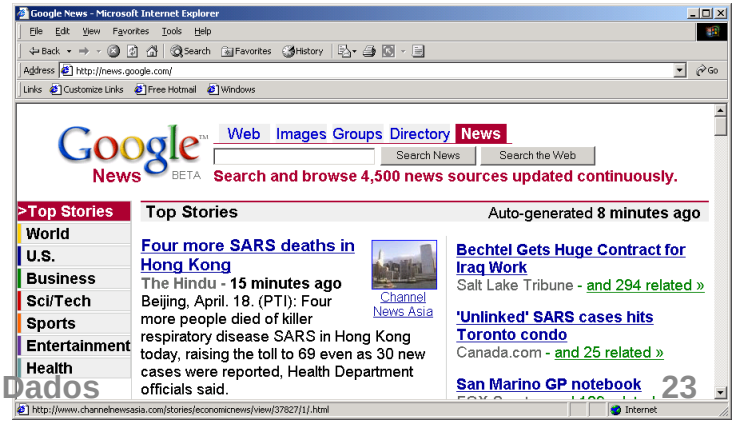




Fonte: Michael Eisen



Uso de K-medias para particionar dados da Sea Surface Temperature (SST) e Net Primary Production (NPP) em clusters que refletem os hemisférios norte e sul.



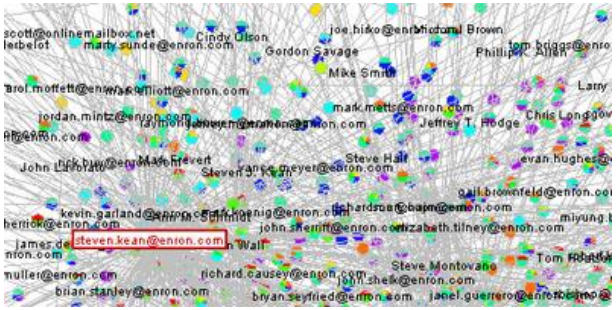
Clustering: aplicação 1

- Segmentação de mercado:
  - **Objetivo:** dividir um mercado em subconjuntos distintos de clientes onde qualquer subconjunto pode ser selecionado como um objetivo de mercado a ser alcançado com uma proposta.
  - **Abordagem:**
    - Colete diferentes atributos de clientes com base em suas informações geográficas e de estilo de vida relacionadas.
    - Encontre clusters de clientes semelhantes.
    - Meça a qualidade de clustering observando padrões de compra de clientes no mesmo cluster versus aqueles de clusters diferentes.

Clustering: aplicação 2

- Clustering de documentos:
  - **Objetivo:** encontrar grupos de documentos que são semelhantes com base nos termos importantes que aparecem neles.
  - **Abordagem:** Identificar os termos que ocorrem com frequência em cada documento. Formar uma medida de similaridade com base nas frequências de diferentes termos. Use-o para agrupar em cluster.

Enron email dataset



Descoberta de regras de associação: definição

- Dado um conjunto de registros cada um dos quais contêm algum número de itens de uma determinada coleção
  - Produzir regras de dependência que preveem a ocorrência de um item com base em ocorrências de outros itens.

TID	*Itens*
1	Pão, Coca, Leite
2	Cerveja, Pão
3	Cerveja, Coca, Fralda, Leite
4	Cerveja, Pão, Fralda, Leite
5	Coca, Fralda, Leite

Regras descobertas:

{Leite} --> {Coca}

{Frada, Leite} --> {Cerveja}

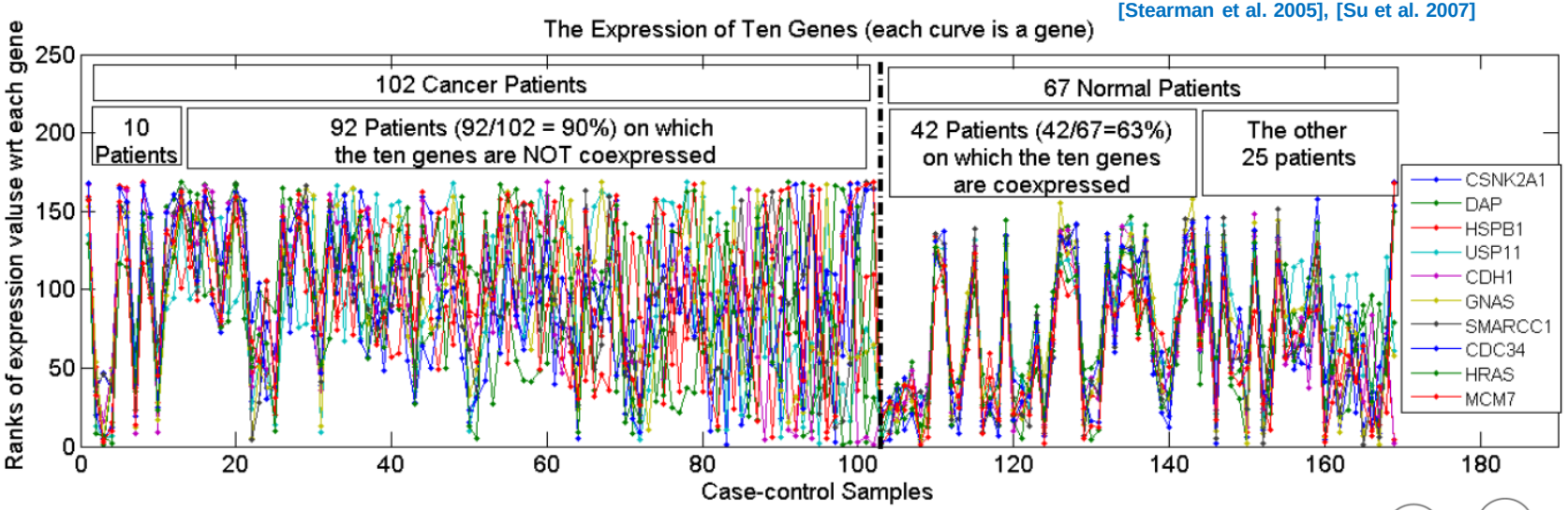
Análise de associação: aplicações

- Análise de mercado
  - As regras são usadas para promoção de vendas, gerenciamento de prateleira e gerenciamento de estoques
- Diagnóstico do alarme da telecomunicação
  - As regras são usadas para encontrar a combinação de alarmes que ocorrem junto frequentemente no mesmo período de tempo
- Informática médica
  - As regras são usadas para encontrar a combinação de sintomas do paciente e resultados do teste associados com determinadas doenças

Análise de associação: aplicações

- Um exemplo de padrão de expressão diferencial subespacial do conjunto de dados de câncer de pulmão

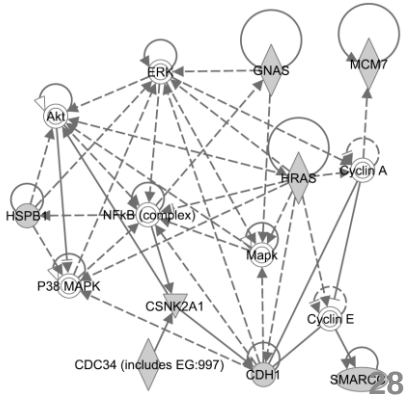
Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]



Enriquecido com a via de sinalização TNF/NFB

que é bem conhecida por estar relacionada ao câncer de pulmão

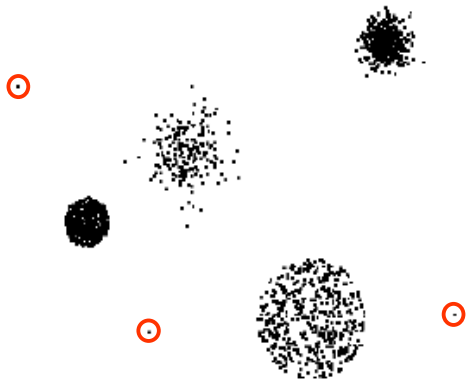
P-valor: 1.4*10-5 (6/10 sobreposição com o caminho)

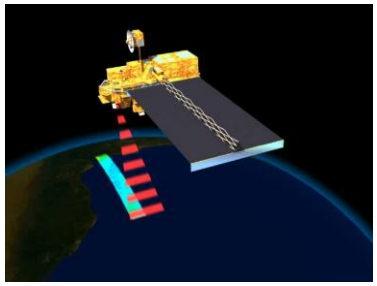


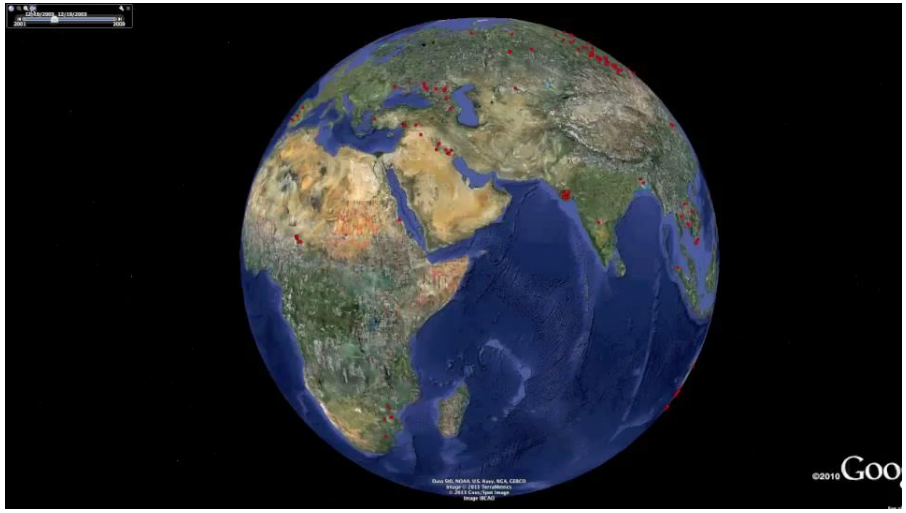
[Fang et al PSB 2010]

Deteção de desvio/anomalia/alteração

- Detectar desvios significativos do comportamento normal
- Aplicações:
  - Detecção de fraude de cartão de crédito
  - Detecção de intrusão na rede
  - Identifique o comportamento anômalo de redes de sensores para monitoramento e vigilância.
  - Detectando alterações na cobertura florestal global.







Desafios Futuros

- Escalabilidade
- Alta dimensionalidade
- Dados heterogêneos e complexos
- Propriedade e distribuição de dados
- Análise não tradicional