MINERAÇÃO DE DADOS
INTRODUÇÃO

Prof. Dr. Rooney R. A. Coelho

PUC-SP

EMENTA

- Abordagem dos Conceitos de mineração de dados, análise exploratória e análise preditiva, Agrupamentos, Associações.

INSTRUMENTOS E CRITÉRIOS DE AVALIAÇÃO

- Pelo menos 75% de presença
- média final deve ser igual ou superior a 5,0 (cinco)

Média Final (MF)

$$MF = \frac{N_1 + N_2}{2} \qquad N_i = \frac{P_i + A_i}{2}$$

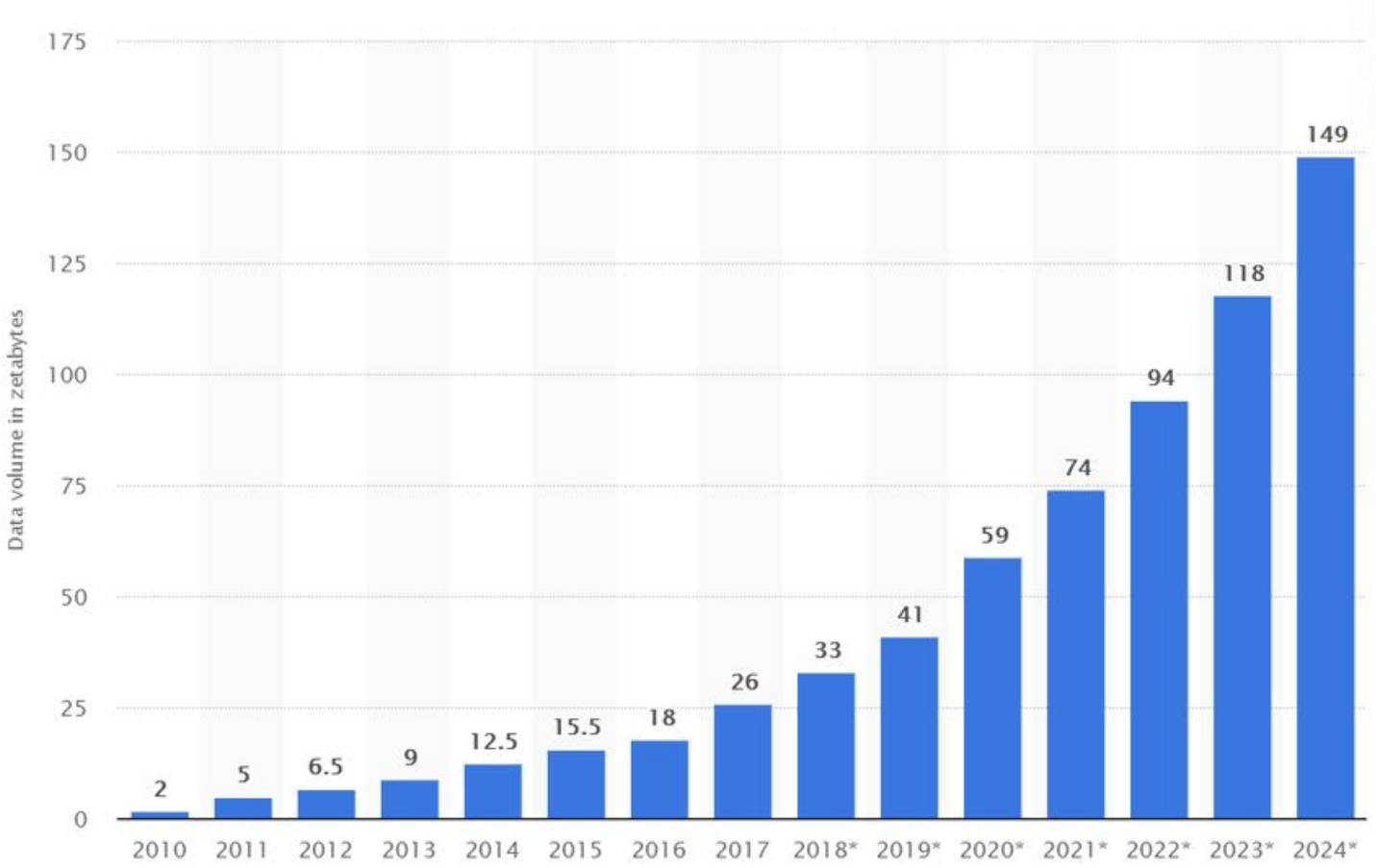
Em que,

- P_i: nota da projeto do bimestre
- A_i: nota de atividades do bimestre

FERRAMENTAS COMPUTACIONAIS

MOTIVAÇÃO

Volume de informação criada, capturada e consumida no mundo



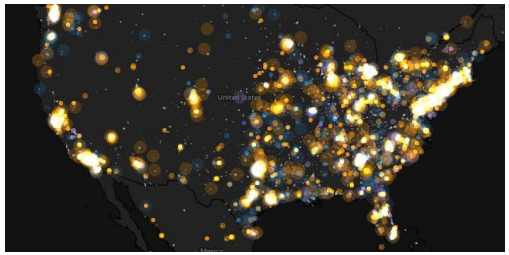
OS DADOS EM GRANDE ESCALA ESTÃO EM TODO LUGAR!

- 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
- O novo paradigma
  - 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.

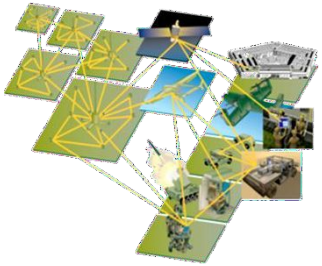
Sistemas de segurança

Comércio Eletrônico

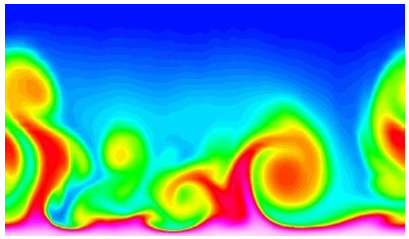
Padrões de transito



Rede Social: Twitter



Redes de sensores



Simulações computacionais

PORQUE A MINERAÇÃO DE DADOS? PONTO DE VISTA COMERCIAL

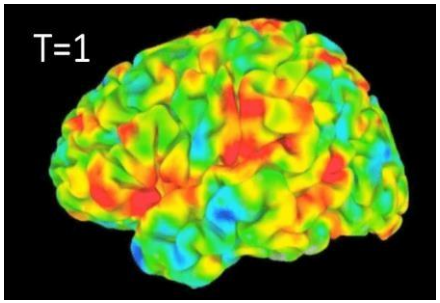
- Uma quantidade enorme de dados é coletada e armazenada
  - Dados da Web
  - Compras nas lojas e supermercados, e-commerce
  - transações bancarias e de cartões de crédito
- A pressão de competição é forte
  - Fornecer serviços melhores e mais customizados (e.g. na gestão de relacionamento com o cliente)

POR QUE A MINERAÇÃO DE DADOS? PONTO DE VISTA CIENTÍFICO

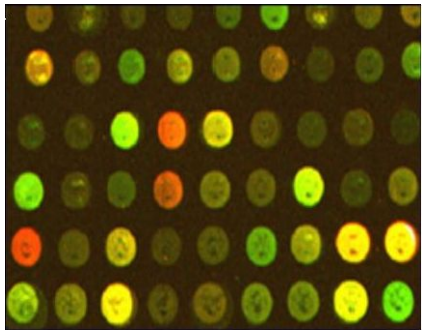
- Sensores remotos em satélites
- Telescópios
- Dados biológicos a altas taxas de transferências
- Simulações científicas

Mineração de dados
- Análise automatizada de grandes conjuntos de dados
- Formação de hipóteses

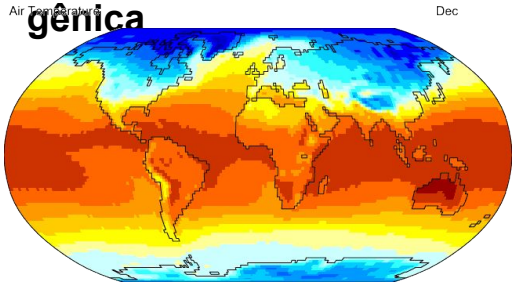
Sky Survey Data



Dados capturados do cérebro



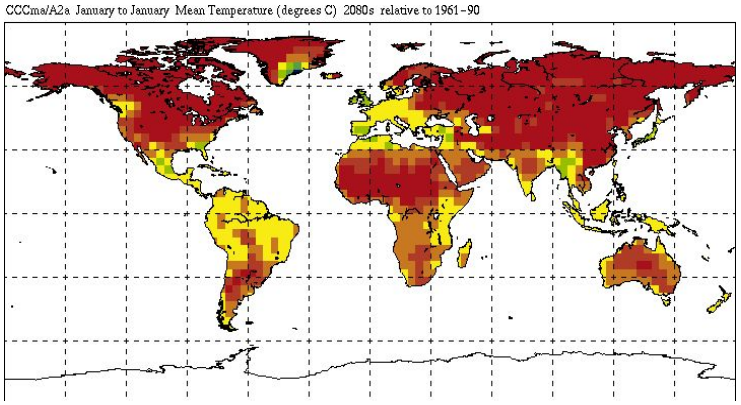
Dados da expressão gênica



Temperatura da superfície da Terra

GRANDES OPORTUNIDADES PARA RESOLVER OS PRINCIPAIS PROBLEMAS DA SOCIEDADE

Melhorar saúde e reduzir custos



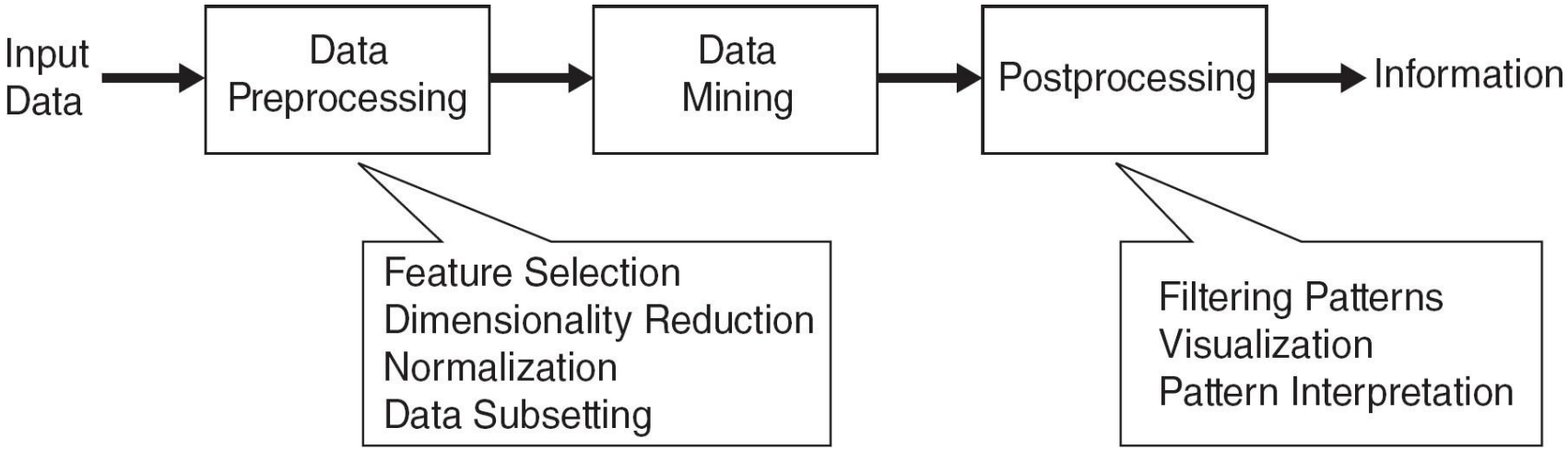
Prever o impacto das alterações climáticas

Encontrar fontes de energia alternativas/verdes

Aumentar a produção agrícola para reduzir a fome e a pobreza

O QUE É A MINERAÇÃO DE DADOS?

- 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



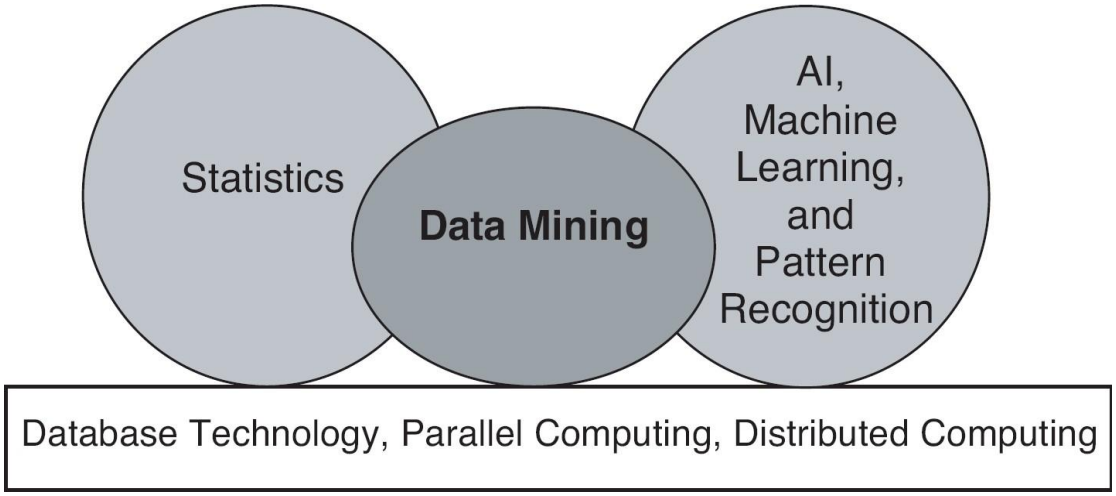
O QUE (NÃO) É A MINERAÇÃO DE DADOS?

- **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)

ORIGENS DA MINERAÇÃO DE DADOS

Empresta ideias de aprendizado de máquina/AI, reconhecimento de padrões, estatísticas e sistemas de banco de dados



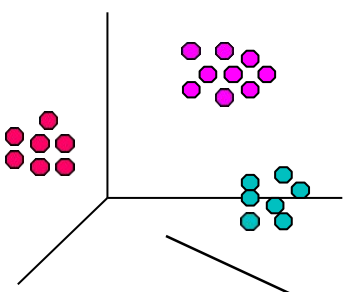
Um componente chave do campo emergente da ciência de dados e da descoberta orientada por dados

Tarefas de mineração de dados

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

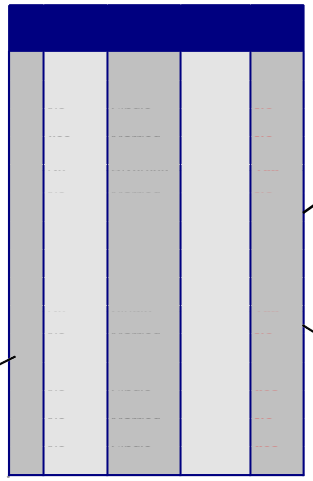
- Métodos Descritivos
  - Encontre padrões interpretáveis por humanos que descrevem os dados.

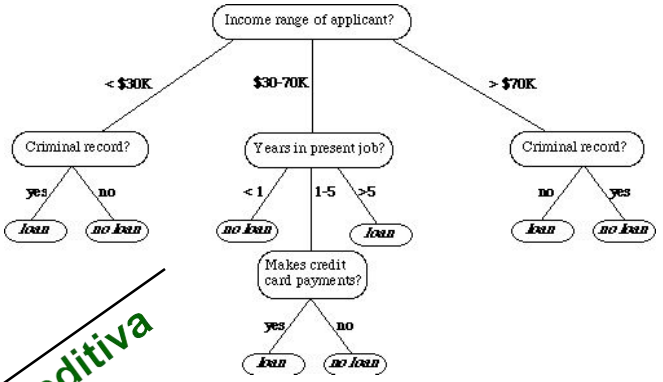
TAREFAS DE MINERAÇÃO DE DADOS



Clustering

Dados

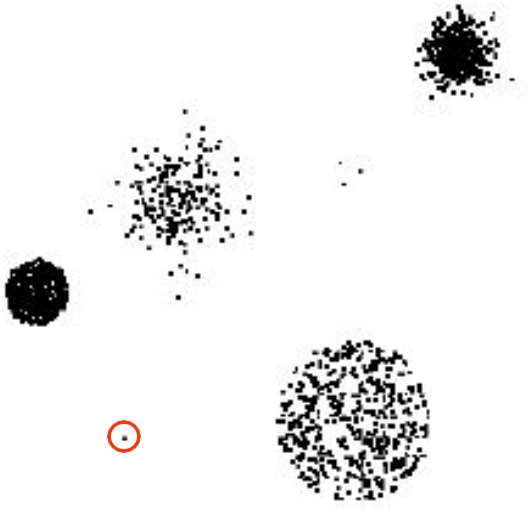




Modelagem Preditiva

Regras de associação

Deteção de anomalias



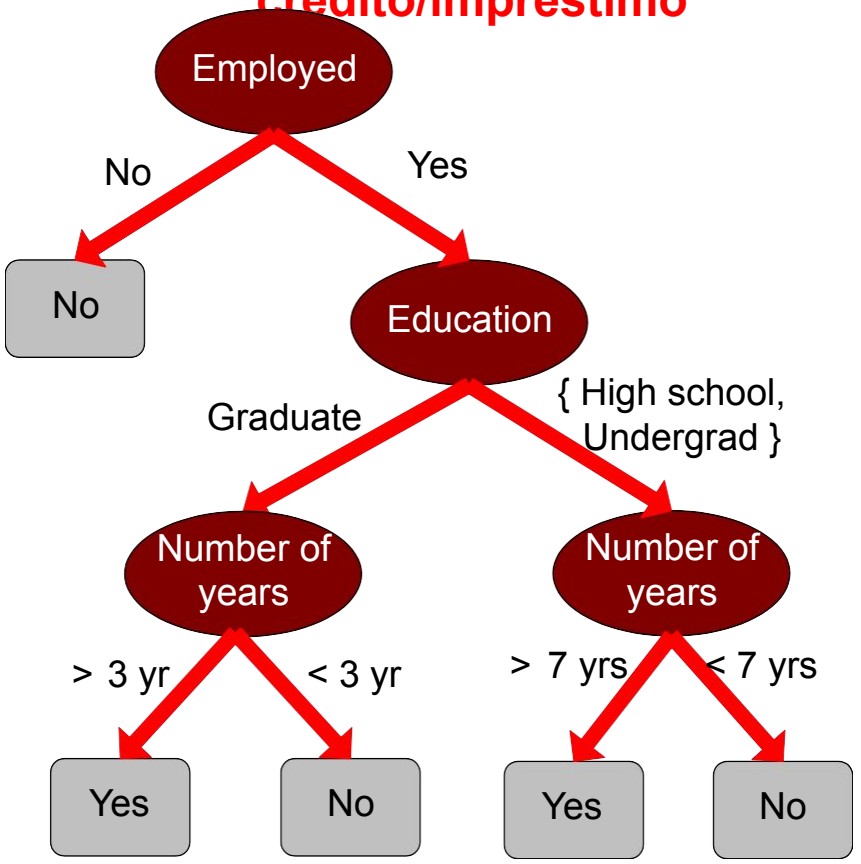
MODELAGEM PREDITIVA: CLASSIFICAÇÃO

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

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
...	...	...	...	...

Modelo de aplicação de crédito/imprestimo

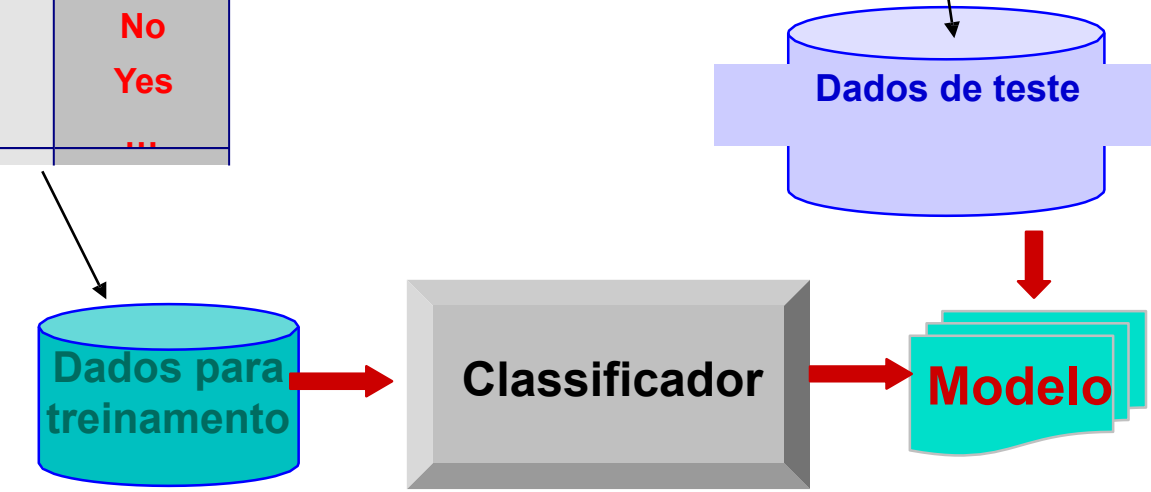


EXEMPLO DE CLASSIFICAÇÃO

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	?
…	…	…	…	…



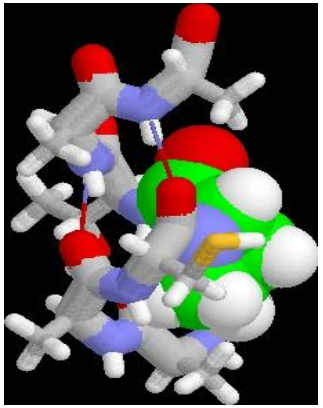
EXEMPLOS DE TAREFAS DE CLASSIFICAÇÃO

- 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.

- Classificação de células tumorais como benignas ou malignas



- Classificação de estruturas de proteínas

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.

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.
  - **Abordagem:**
    - Segmentar a imagem.
    - Medir atributos de imagem (propriedades) - 40 por objeto.
    - Modelar a classe com base nesses atributos.

REGRESSÃO

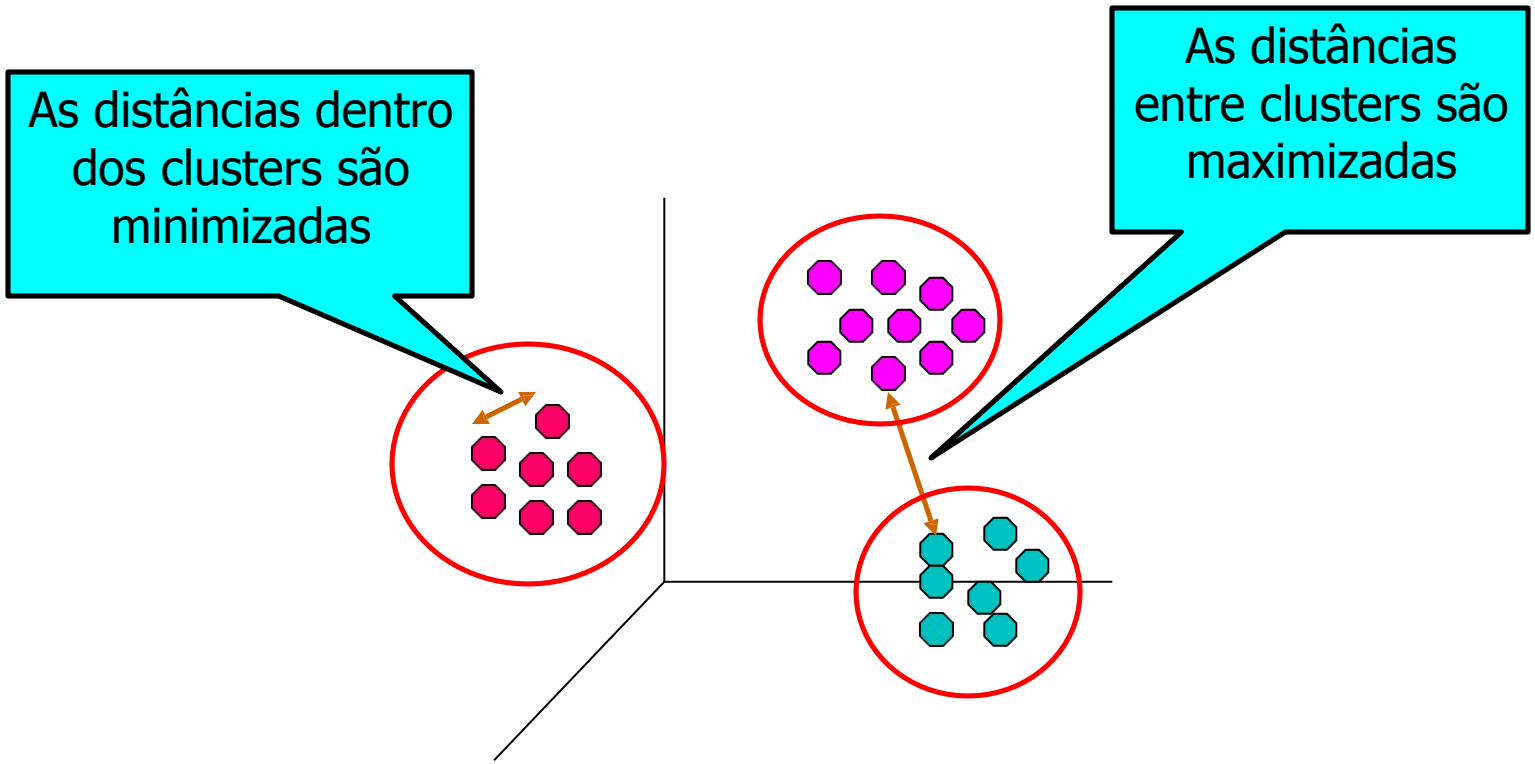
•Prever um valor de uma determinada variável com valor contínuo com base nos valores de outras variáveis, assumindo um modelo linear ou não linear de dependência.

•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



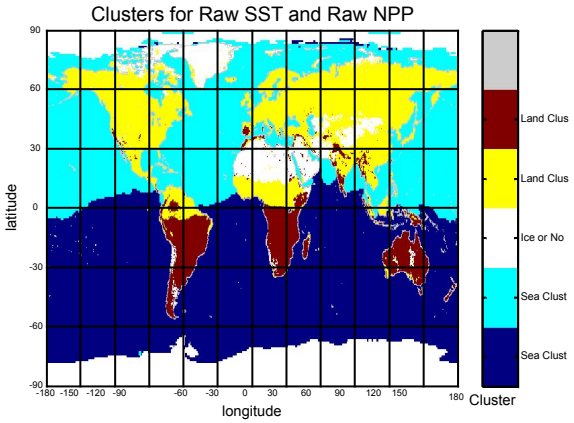
APLICAÇÕES DE ANÁLISE DE CLUSTERS

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



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.

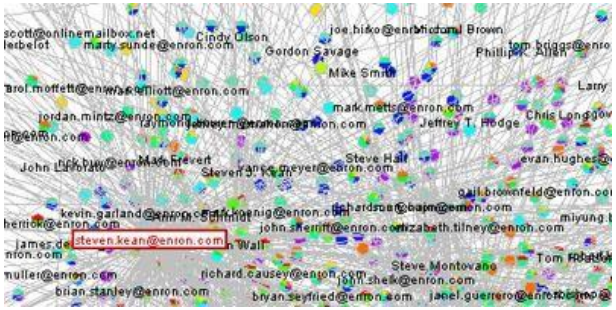
CLUSTERING: APLICAÇÃO 2

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

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.

TIPOS DE DADOS

INSTÂNCIAS DE DADOS

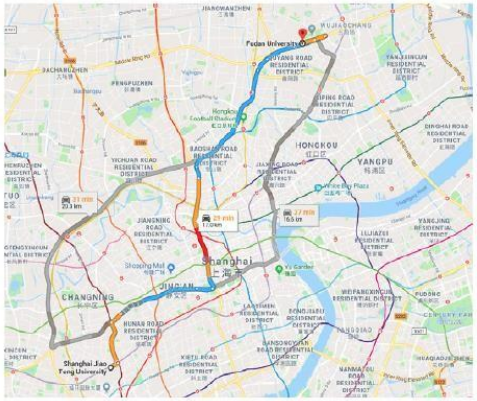
Um artigo de notícias

Uma imagem

Uma música

Perfil de usuário no Facebook

Um histórico de aluno



Uma trajetória no Mapa Google

ATRIBUTOS DE DADOS

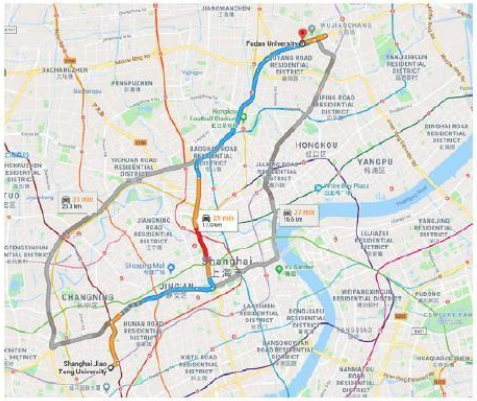
O número de ocorrências de palavra 'Brasil' em um artigo de notícias

O valor RGB do primeiro pixel na primeira linha do lado esquerdo

O número de ocorrências da nota G#

Um conjunto de amigos de um usuário do Facebook

A nota de Cálculo no histórico de aluno



O posição e tempo do terceiro ponto da trajetória

O QUE SÃO DADOS?

- Coleção de ***objetos de dados*** e seus ***atributos***
- Um ***atributo*** é uma propriedade ou característica de um objeto
  - Exemplos: cor dos olhos de uma pessoa, temperatura, etc.
  - Atributo também é conhecido como variável, campo, característica, dimensão ou recurso
- Uma coleção de atributos descrevem um ***objeto***
  - Objeto também é conhecido como registro, ponto, caso, exemplo, entidade ou instância

Atributos

Objetos

Tid	Reembolso	Estado Civil	Rendimento tributável	Fraude
1	Sim	Solteiro	125K	No
2	Não	Casado	100K	No
3	Não	Solteiro	70K	No
4	Sim	Casado	120K	No
5	Não	Divorciado	95K	Yes
6	Não	Casado	60K	No
7	Sim	Divorciado	220K	No
8	Não	Solteiro	85K	Yes

TIPOS DE ATRIBUTOS

- Existem diferentes tipos de atributos
  - Nominal
    - Exemplos: números de identificação, cores dos olhos, códigos postais
  - Ordinal
    - Exemplos: classificações (por exemplo, sabor de batatas fritas em uma escala de 1 a 10), notas, altura {alto, médio, baixo}
  - Intervalo
    - Exemplos: datas do calendário, temperaturas em Celsius ou Fahrenheit.
  - Proporção
    - Exemplos: temperatura em Kelvin, comprimento, tempo, contagens

TIPOS DE ATRIBUTOS

Nominal

- Valor do atributo é um nome para algo
- Uma categoria, um código, um estado
- Exemplos: estado civil, cor do cabelo, ocupação
- Podem ser representados por números arbitrários
  - Mas não tem sentido efetuar operações entre estes valores,

  não são quatitativos, nem tem ordenação
- Não há média nem mediana

TIPOS DE ATRIBUTOS

Binário

- Atributos com dois valores: 0 ou 1
- Ausente ou presente, sim ou não
- Exemplo: fumante?, possui carro?
- Simétrico: ambos valores ão relevantes
- Assimétrico: um valor é mais relevante (normalmente, o valor 1 é utilizado)

TIPOS DE ATRIBUTOS

Ordinal

- Valorem possuem uma ordem (ranking)
- O valor em si não tem significado
- Exemplo: notas, tamanho P-M-G-XG, escala de satisfação
- Podem vir da discretização de quantidades numéricas

TIPOS DE DADOS

Numérico

- Quantitativo, quantidade mensurável
- Escala por intervalo (interval-scaled):
  - Escala de unidades de mesmo tamanho
  - Ordem, há diferença entre valores
  - Não há zero verdadeiro, indicando ausência
  - Exemplo: temperatura em Celsius, dias de calendário
- Escala por razão (ratio-scaled):
  - Há zero verdadeiro
  - Exemplo: temperatura em Kelvin, valor monetário

ATRIBUTOS DISCRETOS E CONTÍNUOS

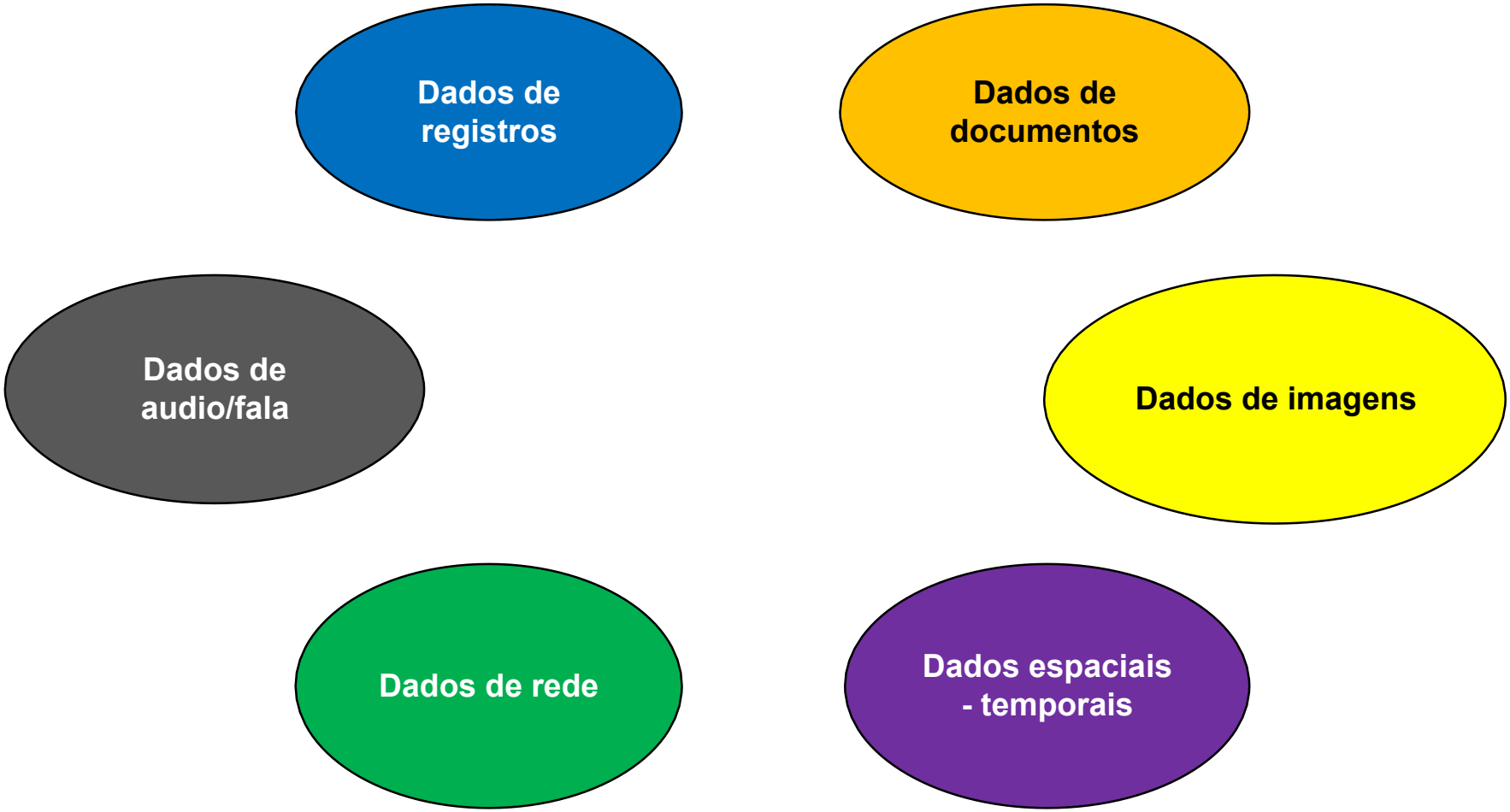
Atributo discreto

- Há apenas um conjunto finito ou contável e infinito de valores
- Exemplos: códigos postais, profissão, contagens ou conjunto de palavras em uma coleção de documentos
- Normalmente representado como variável inteira.
- Atenção: atributos binários são um caso especial de atributos discretos

Atributo contínuo

- Tem números reais como valores de atributo
- Exemplos: temperatura, altura ou peso.
- Praticamente, valores reais só podem ser medidos e representados usando um número finito de dígitos.
- Atributos contínuos são normalmente representados como variáveis de ponto flutuante.

6 GRANDES TIPOS DE DADOS



DADOS DE REGISTRO

Muito comum em bancos de dados:

- Cada linha representa uma instância de dados
- Cada coluna representa um atributo

WEEKDAY	GENDER	AGE	CITY
Tuesday	Male	28	London
Monday	Female	24	New York
Tuesday	Female	36	Hong Kong
Thursday	Male	17	Tokyo

JSON Format:

```
{
    WEEKDAY: Monday;
    GENDER: Female;
    AGE: 24;
    CITY: New York;
}
```

DADOS DE REGISTRO

Dados que consistem uma coleção de registros, cada um deles consiste um conjunto fixo de atributos.

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

MATRIZ DE DADOS

Conjunto fixo de atributos numéricos, pontos em um espaço multidimensional, onde cada dimensão representará um atributo distinto

Esse conjunto de dados pode ser representado por uma matriz *m* por *n*, onde há *m* linhas, uma para cada objeto e *n* colunas, uma para cada atributo

Projection of x Load	**Projection of y load**	**Distance**	**Load**	**Thickness**
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

DADOS DE DOCUMENTOS

Uma sequência de palavras/fichas que representa significados semânticos de humano.

- A mineração de texto, também conhecida como mineração de dados de texto, aproximadamente equivalente à análise de texto, é o processo de derivar informações de alta qualidade do texto.

Bag-of-Words Format:

```
{
    text: 4;
    mining: 2;
    also: 1;
    referred: 1;
    to: 2;
    as: 1;
    data: 1;
    roughly: 1;
    equivalent: 1;
    analytics: 1;
    is: 1;
    the: 1;
    process: 1;
    of: 1;
    deriving: 1;
    high-quality: 1;
    information: 1;
    from: 1;
}
```

DADOS DE DOCUMENTOS

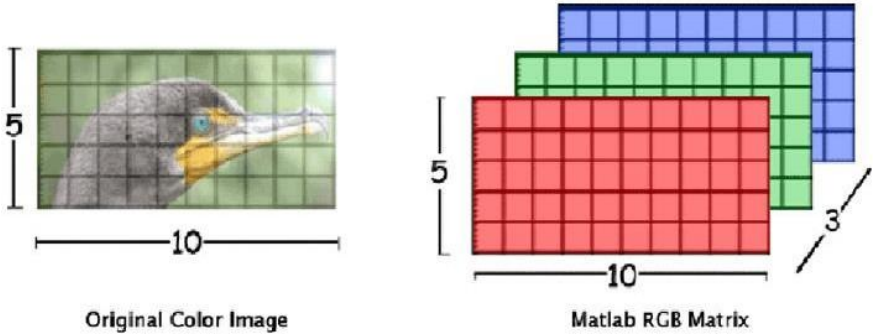
Cada documento se torna um vetor de 'termo'

- Cada termo é um componente (atributo) do vetor
- O valor de cada componente é o número de ocorrências do termo no documento.

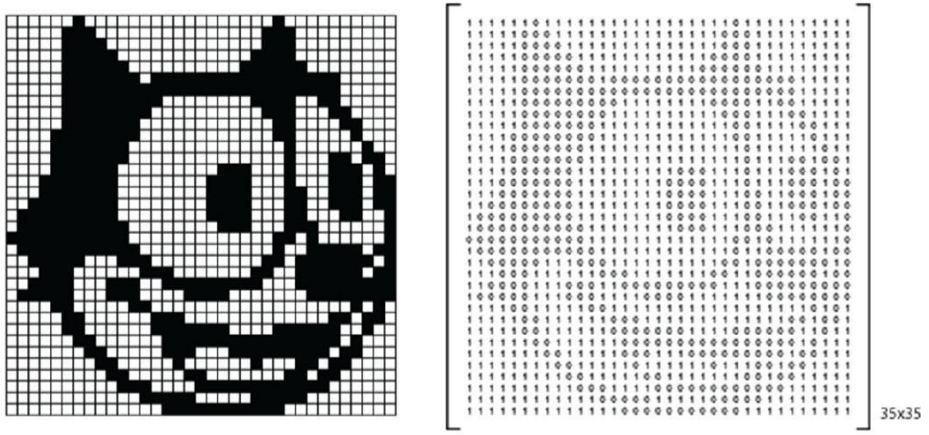
	team	coach	play	ball	score	game	win	lost		timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2	
Document 2	0	7	0	2	1	0	0	3	0	0	
Document 3	0	1	0	0	1	2	2	0	3	0	

DADOS DE IMAGENS

Uma matriz de 3 camadas (3*altura*largura) de valores entre [0,255]



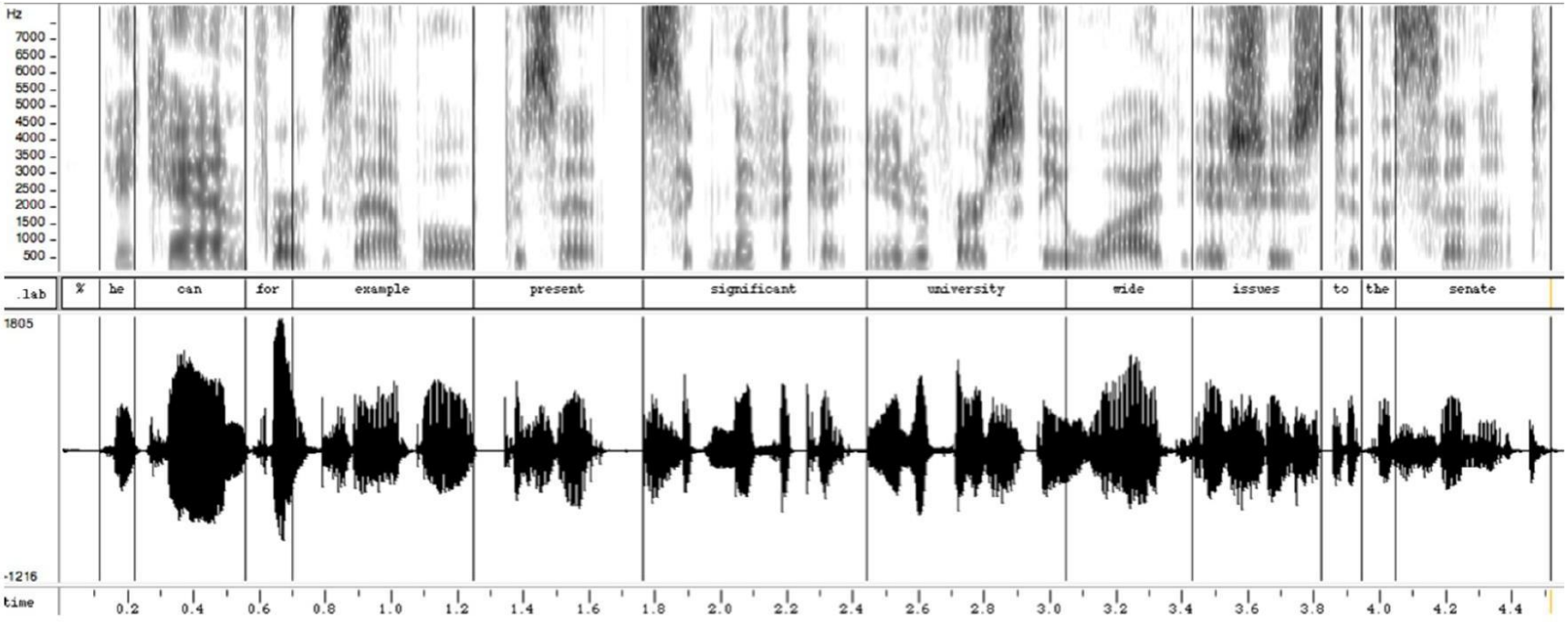
Um caso simples: imagem binária



DADOS DE ÁUDIO

Uma sequência de vetores reais multidimensionais

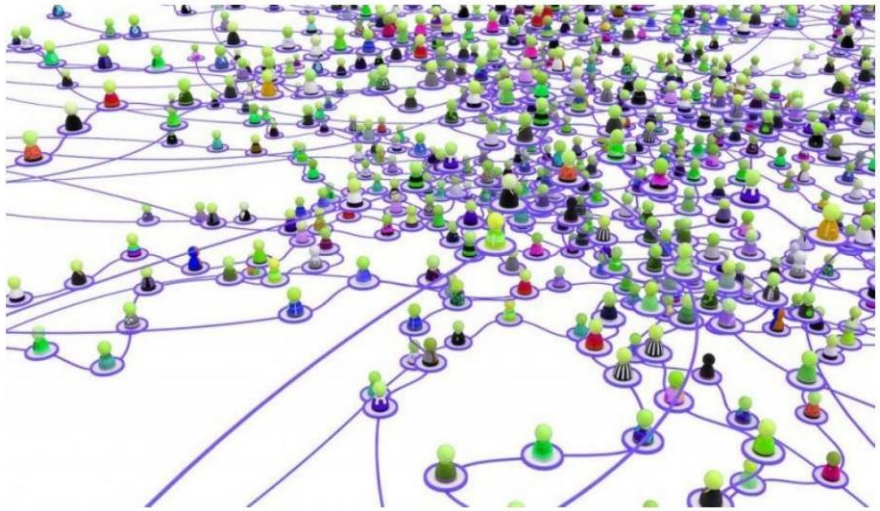
- Decodificação direta dos dados de áudio / fala



DADOS DE REDE

Um grafo direcionado / não direcionado

- Possivelmente com informações adicionais para nós e arestas

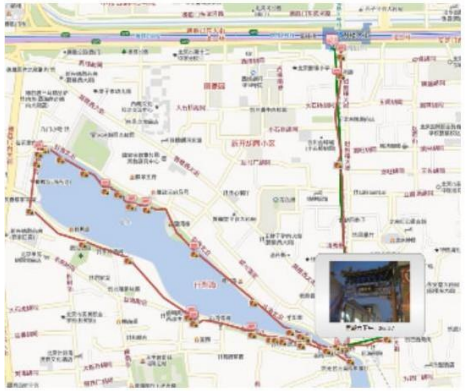


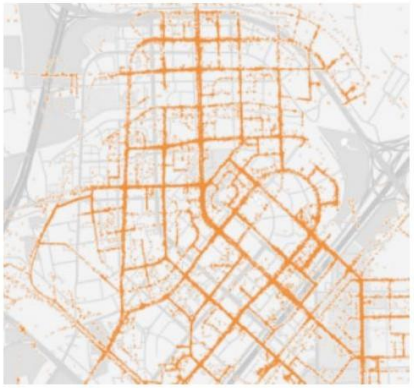
Friendship Format:

Alice	Bob
Bob	Carl
Carl	Victor
Bob	Victor
Alice	Victor
…	

DADOS ESPACIAIS E TEMPORAIS

Uma sequência de tuplas (hora, local, informações)

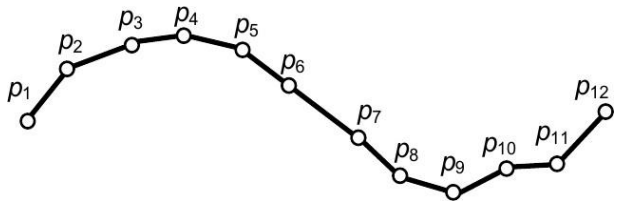




Uma trajetória espaço-temporal

$$p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$$

$$p_i = (t, x, y, a)$$



DADOS ORDENADOS

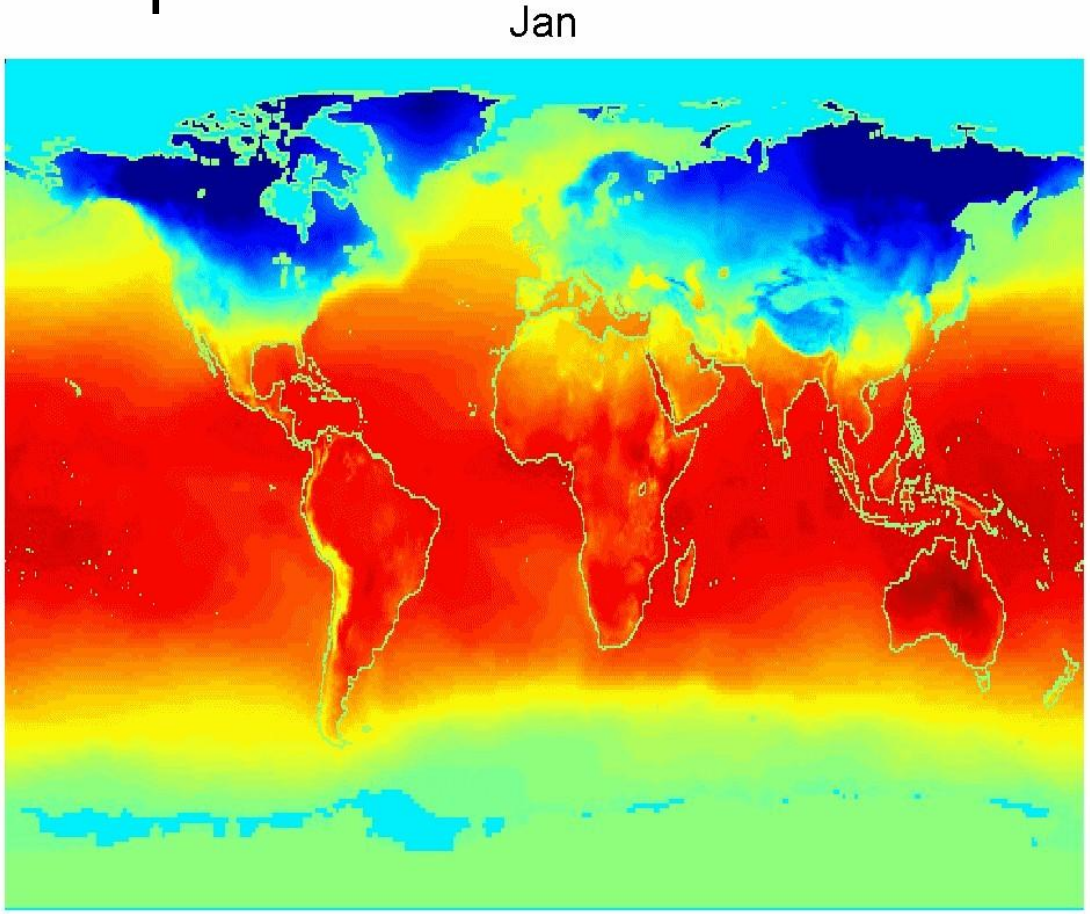
Dados de sequência genômica

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

DADOS ORDENADOS

Dados espaciais e temporais

Temperatura média mensal da terra e do oceano



QUALIDADE DOS DADOS

A baixa qualidade dos dados afeta negativamente muitos esforços de processamento de dados

Exemplo de mineração de dados: um modelo de classificação para detectar pessoas que são riscos de empréstimo é criado usando dados ruins

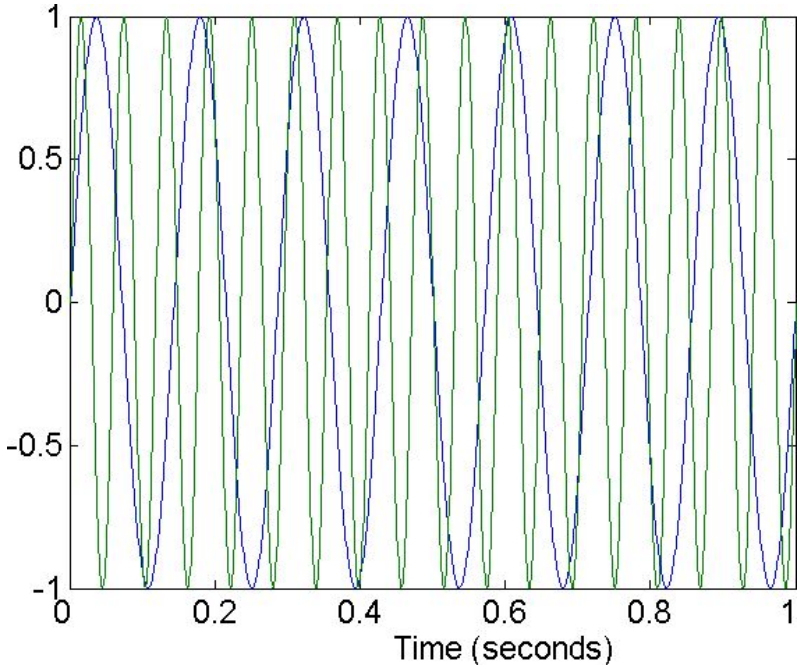
- Empréstimos para alguns candidatos são negados, mesmo que eles tem condições de devolve-los
- Mais empréstimos são concedidos aos indivíduos que não tem condições de pagamento

RUÍDO

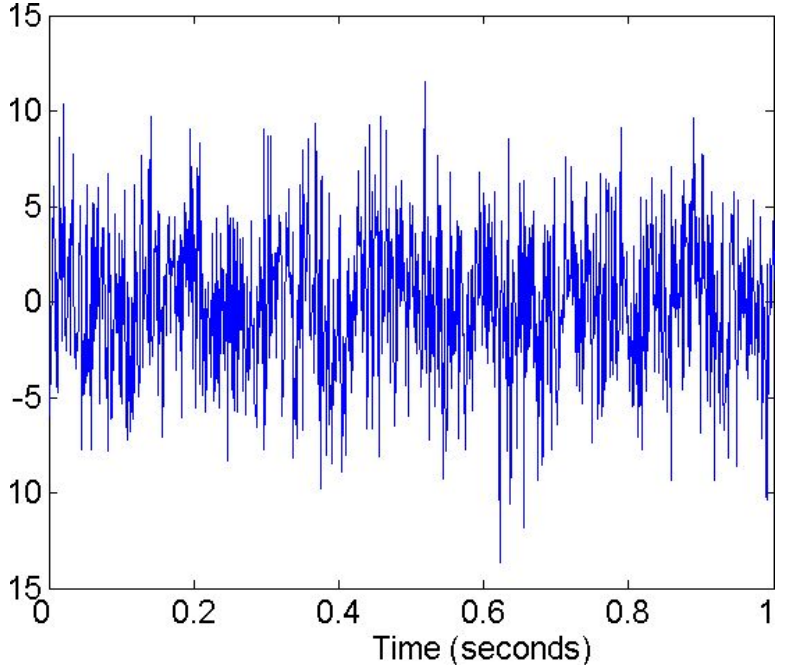
Para objetos, o ruído é um objeto estranho

Para atributos, o ruído refere-se à modificação dos valores originais

- Exemplos: distorção da voz de uma pessoa ao falar em um telefone pobre e "neve" na tela da televisão



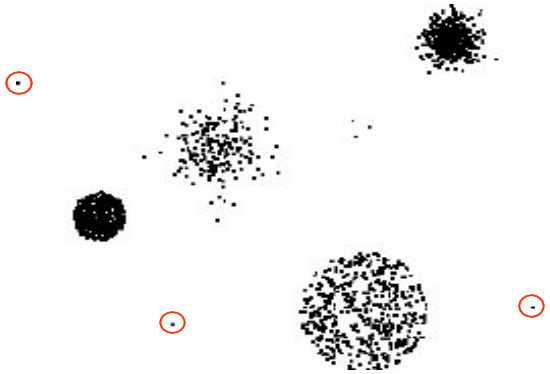
Duas ondas senoidais



Duas ondas senoidais + Ruído

OUTLIERS

- ***Outliers*** são objetos de dados com características que são consideravelmente diferentes da maioria dos outros objetos de dados no conjunto de dados
  - **Case 1:** Outliers são ruídos que interferem com análise de dados
  - **Case 2:** Outliers são os objetivos de nossa análise
    - Fraude de cartão de crédito
    - Detecção de intrusão



VALORES AUSENTES

Razões para valores ausentes

- As informações não foram coletadas

  (por exemplo, as pessoas não querem informar suas idades ou pesos)
- Os atributos não são aplicáveis aos todos os casos (por exemplo, o rendimento anual não é aplicável às crianças)

O que fazer com valores ausentes

- Eliminar objetos de dados ou variáveis
- Estimar valores ausentes
  - Exemplo: séries temporais de temperatura
  - Exemplo: resultados censitários
- Ignorar o valor ausente durante a análise

DADOS DUPLICADOS

Conjunto de dados pode incluir objetos de dados que são duplicados, ou quase duplicatas um do outro

- Principal problema ao mesclar dados de fontes heterogêneas

Exemplos:

- Mesma pessoa com vários endereços de e-mail

Limpeza de dados

- Processo de lidar com problemas de dados duplicados

Quando os dados duplicados não devem ser removidos?