ACH5504 – Mineração de Dados

Notas de aulas baseadas no livro

"*Introduction to Data Mining*"

Tan, Steinbach, Karpatne, Kumar

Resumo

- Instâncias de dados, atributos e objetos
- Tipos de dados
- Qualidade dos dados
- Descrição estatística de dados

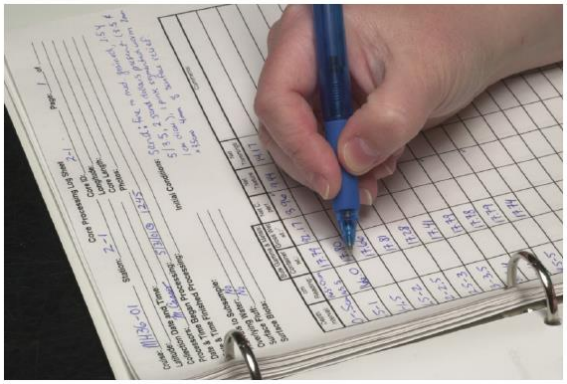
Instâncias de dados

Um artigo de notícias

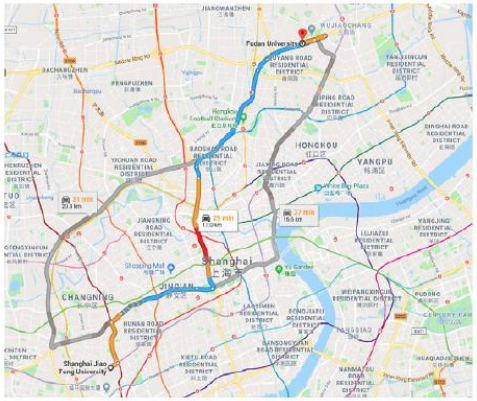
Uma imagem

Uma música

Perfil de usuário no Facebook



Um histórico de aluno



Uma trajetória no Mapa Google

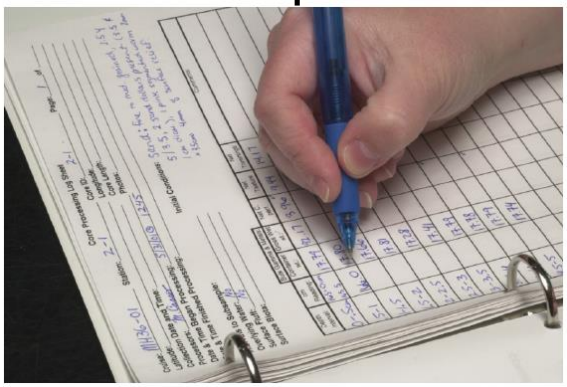
Atributos de dados

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

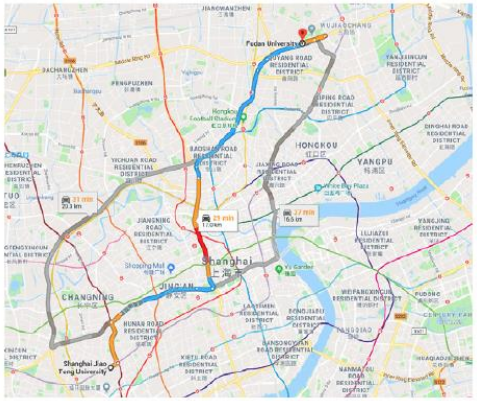
O valor RGB do primeiro pixel na primeira linha do lado esquerdo

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

Uma visão mais completa dos dados

- Os dados podem ter partes
- As diferentes partes dos dados podem ter relacionamentos
- Mais geralmente, os dados podem ter estrutura
- Os dados podem ser incompletos
- Discutiremos isso com mais detalhes depois

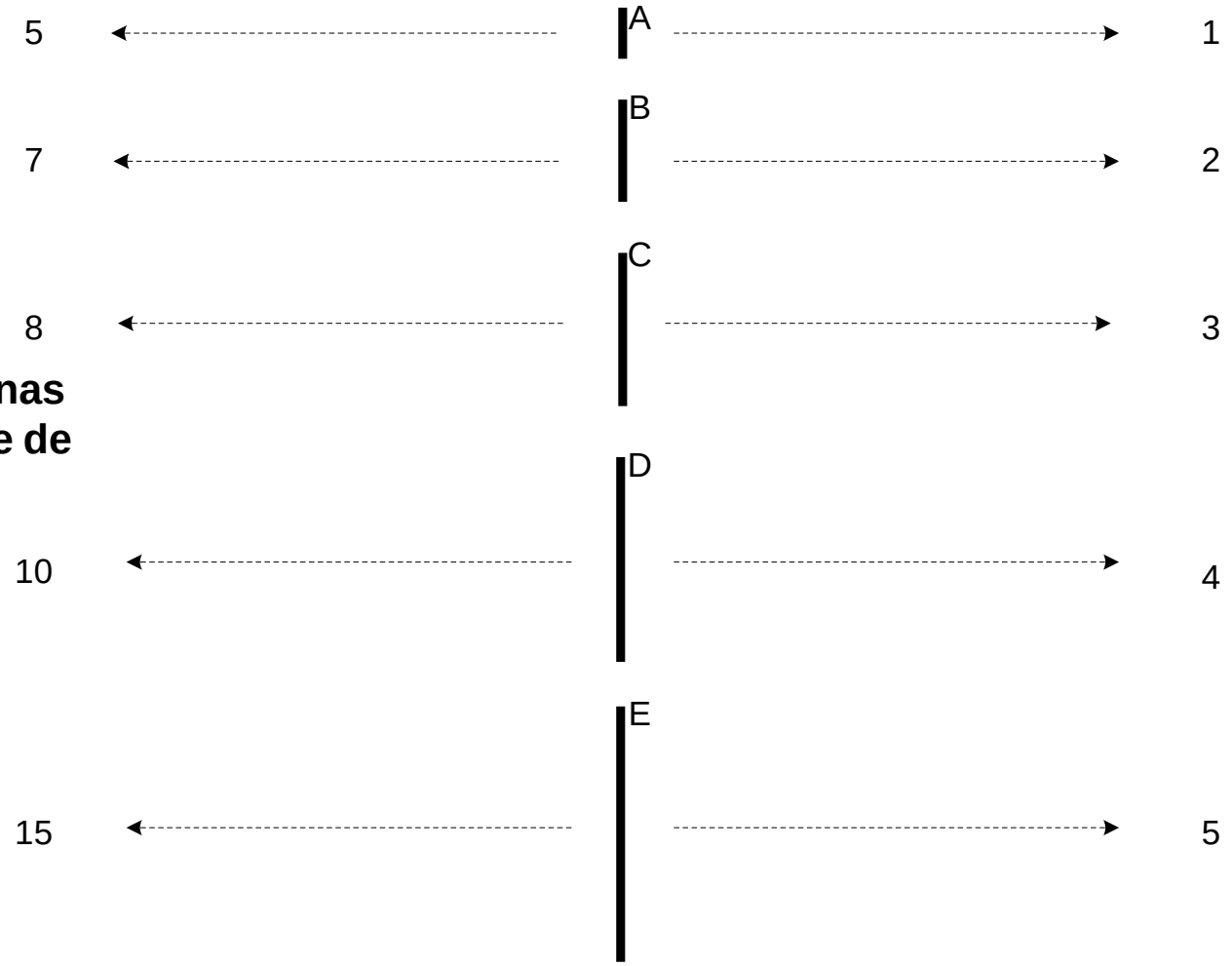
Valores de atributo

- ***Valores de atributo*** são números ou símbolos atribuídos a um atributo para um determinado objeto

- Distinção entre atributos e valores de atributo
  - O mesmo atributo pode ser mapeado para valores de atributo diferentes
    - Exemplo: a altura pode ser medida em pés ou metros
  - Atributos diferentes podem ser mapeados para o mesmo conjunto de valores
    - Exemplo: valores de atributo para ID e idade são números inteiros
    - Mas propriedades de valores de atributo podem ser diferentes

Medição de comprimento

- A maneira como você mede um atributo pode não corresponder as propriedades de atributos.



Esta escala preserva apenas a propriedade de ordenação.

Esta escala preserva as propriedades de ordenação e aditividade.

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
    - Exempl: temperatura em Celsius, dias de calendário
  - Escala por razão (ratio-scaled):
    - Há zero verdadeiro
    - Exemplo: temperatura em Kelvin, valor monetário

Propriedades de valores de atributo

- O tipo de um atributo depende de qual das seguintes propriedades/operações que ele possui:
  - Distinção: $= \neq$
  - Ordem: $< >$
  - As diferenças são significativas: $+ -$
  - As proporções são significativas: $* /$
  - Atributo nominal: distinção
  - Atributo ordinal: distinção & ordem
  - Atributo de intervalo: distinção, ordem & diferenças significativas
  - Atributo de proporção: todas as 4 propriedades/operações

Diferença entre razão e intervalo

- É fisicamente significativo dizer que uma temperatura de 10° é duas vezes a de 5° em
  - a escala de Celsius?
  - a escala de Fahrenheit?
  - a escala de Kelvin?

- Considere medir a altura acima da média
  - Se a altura do João é de três centímetros acima da média e a altura do Marcus é de seis centímetros acima da média, então nós dizemos que o Marcus é duas vezes mais alto que o João?
  - Esta situação é análoga à da temperatura?

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.

Atributos assimétricos

- Somente a presença (um valor de atributo diferente de zero) é considerada como importante
  - Palavras presentes em documentos.
  - Itens presentes em transações de clientes.

- Se alguém encontrar um amigo no supermercado, ia dizer o seguinte?

  "Vejo que nossas compras são parecidas, pois nós não compramos as mesmas coisas similares."

- Precisamos de dois atributos binários assimétricos para representar um atributo binário comum
  - Análise de associação usa atributos assimétrico

- Os atributos assimétricos geralmente surgem de objetos que são conjuntos

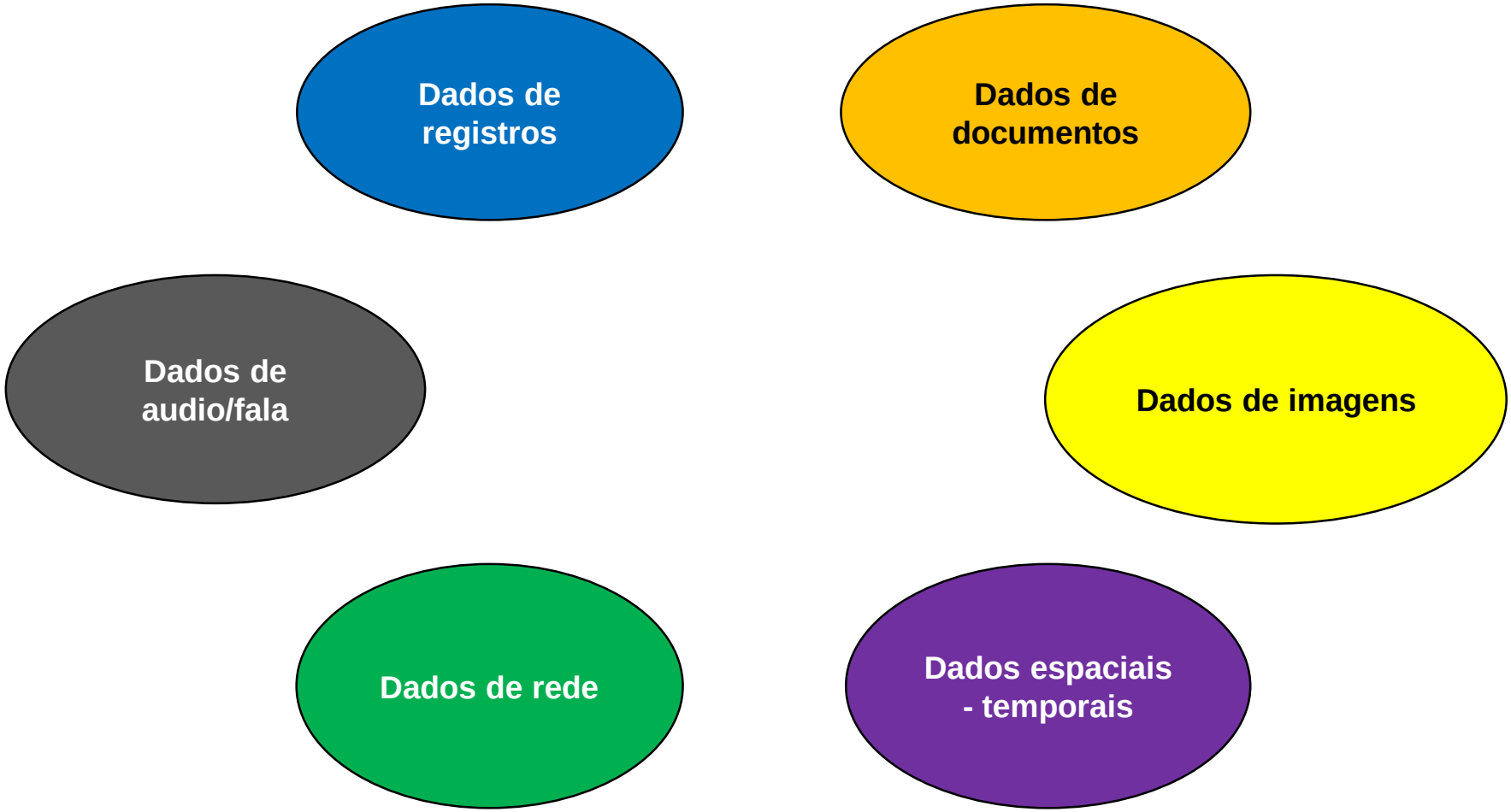
Exemplos mais complicados

- Números de identidade (ID)
  - Nominal, ordinal ou intervalo?
- Número de cilindros em um motor de automóvel
  - Nominal, ordinal ou proporção?
- Escala com tendência
  - Intervalo ou proporção?

Pontos importantes para tipos de atributos

- Os tipos de operações que você escolher devem ser "significativos" para o tipo de dados que você tem
  - Distintividade, ordem, intervalos significativos e proporções significativas são apenas quatro propriedades de dados
  - O tipo de dados que você vê – muitas vezes números ou cadeias de caracteres – não pode capturar todas as propriedades ou pode sugerir propriedades que não estão lá
  - A análise pode depender dessas outras propriedades dos dados
    - Muitas análises estatísticas dependem apenas da distribuição
  - Muitas vezes o que é significativo é medido por significância estatística
  - Mas no final, o que é significativo é medido pelo domínio

6 Majores Tipos de Dados



Tipos de conjuntos de dados

- Registro
  - Matriz de dados
  - Dados de documentos
  - Dados de transações
- Gráfico
  - World Wide Web
  - Estruturas moleculares
- Ordenados
  - Dados espaciais
  - Dados temporais
  - Dados sequenciais
  - Dados de sequências genéticas

Características importantes dos dados

- Dimensionalidade - número de atributos
  - Dados de alta dimensão trazem uma série de desafios
- Esparcidade ou dispersão
  - Somente a presença conta, atributos ausentes ou 0
- Resolução
  - Os padrões dependem da escala
- Tamanho
  - O tipo de análise pode depender do tamanho dos 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

- Se os objetos de dados tiverem o mesmo conjunto fixo de atributos numéricos, os objetos de dados poderão ser pensados como 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

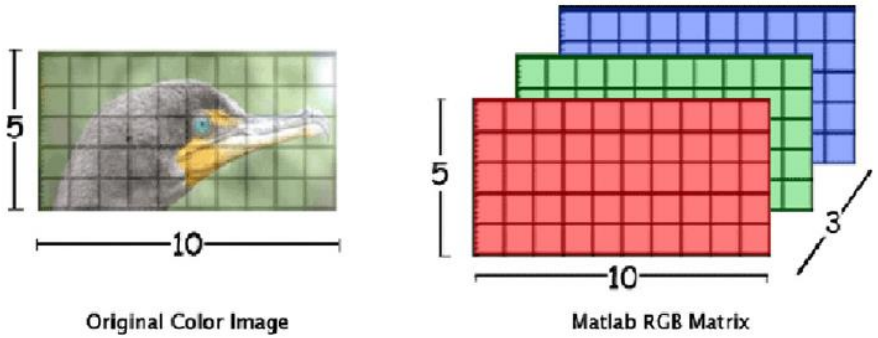
Dados de transações

- Um tipo especial de dados de registro, onde
  - Cada registro (transação) envolve um conjunto de itens.
  - Por exemplo, considere um supermercado. O conjunto de produtos comprados por um cliente durante uma viagem de compras constitui uma transação, enquanto os produtos individuais que foram comprados são os itens.

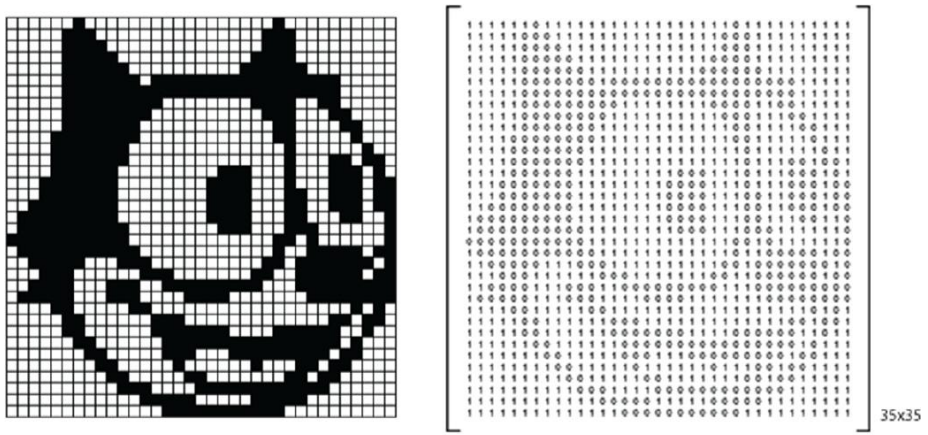
TID	*Itens*
1	Pão, Coca, Leite
2	Cerveja, Pão
3	Cerveja, Coca, Fralda, Leite
4	Cerveja, Pão, Fralda, Leite
5	Coca, Fralda, Leite

Dados de Imagens

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

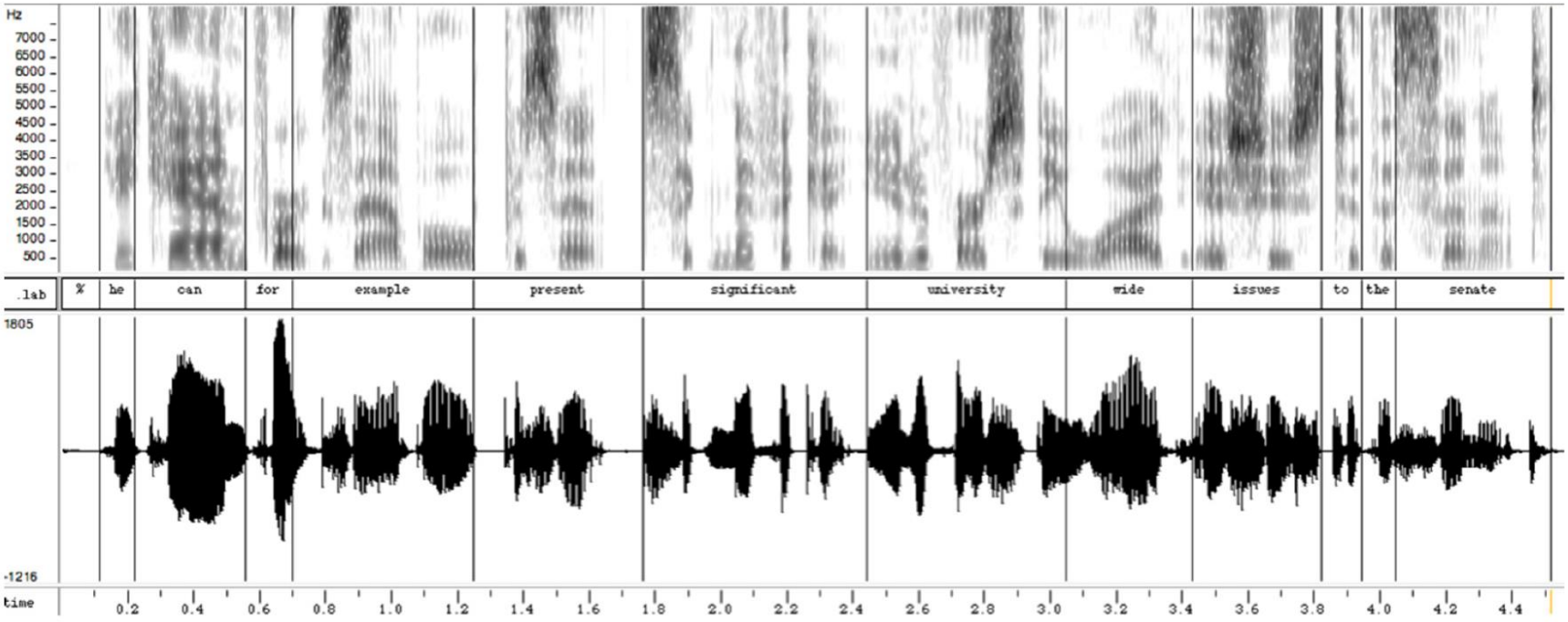


- Um caso simples: imagem binária



Dados de áudio

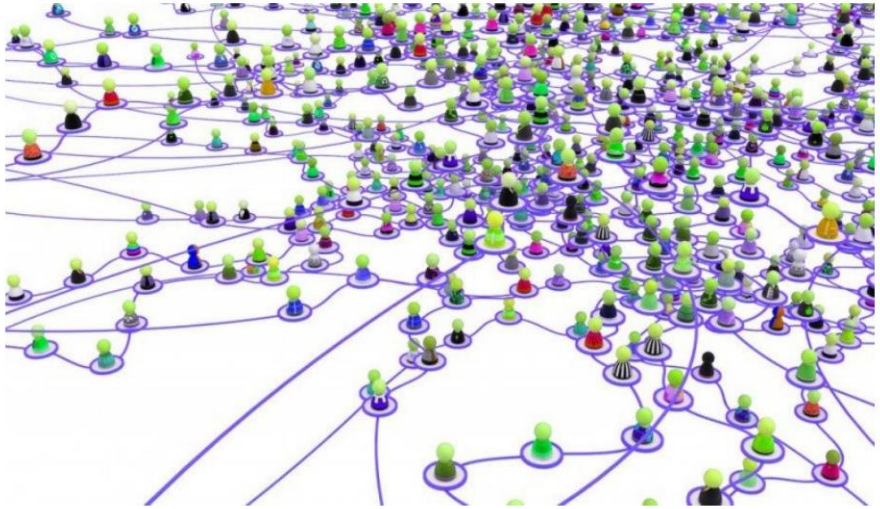
- Uma sequência de vetores reais multidimensionais
  - Decodificação direta dos dados de áudio / fala



http://languagelog.ldc.upenn.edu/nll/?p=8116

Dados de rede

- Um gráfico direcionado / não direcionado
  - Possivelmente com informações adicionais para nós e arestas



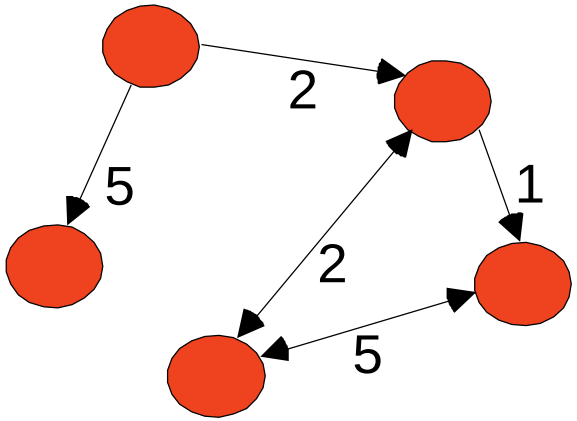
Friendship Format:

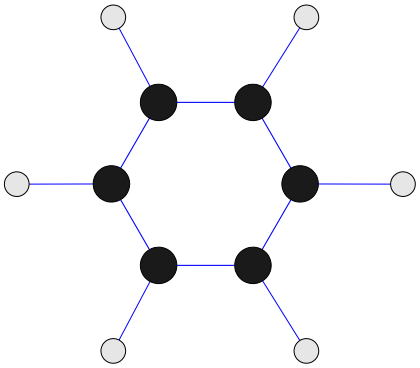
Alice	Bob
Bob	Carl
Carl	Victor
Bob	Victor
Alice	Victor
…	

Stanford network dataset collection: https://snap.stanford.edu/data/

Gráficos de dados

- Exemplos: gráfico genérico, de uma molécula ou páginas da Web





Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography
(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

General Data Mining

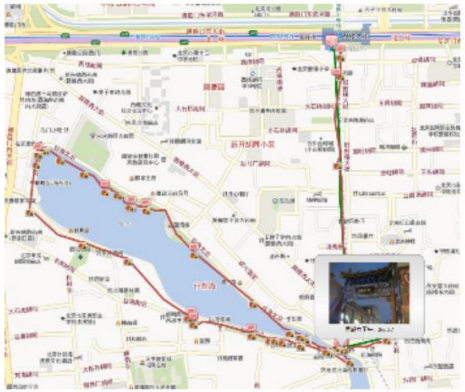
Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

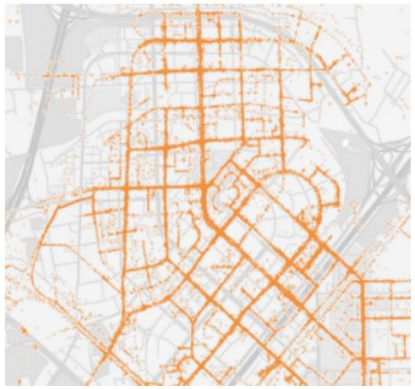
Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Molécula do benzeno: C6H6

Dados espaciais e temporais

- Uma sequência de tuplas (hora, local, informações)

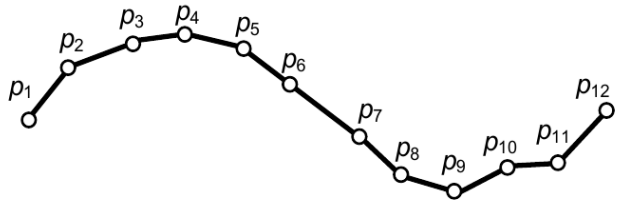




- Uma trajetória espaço-temporal

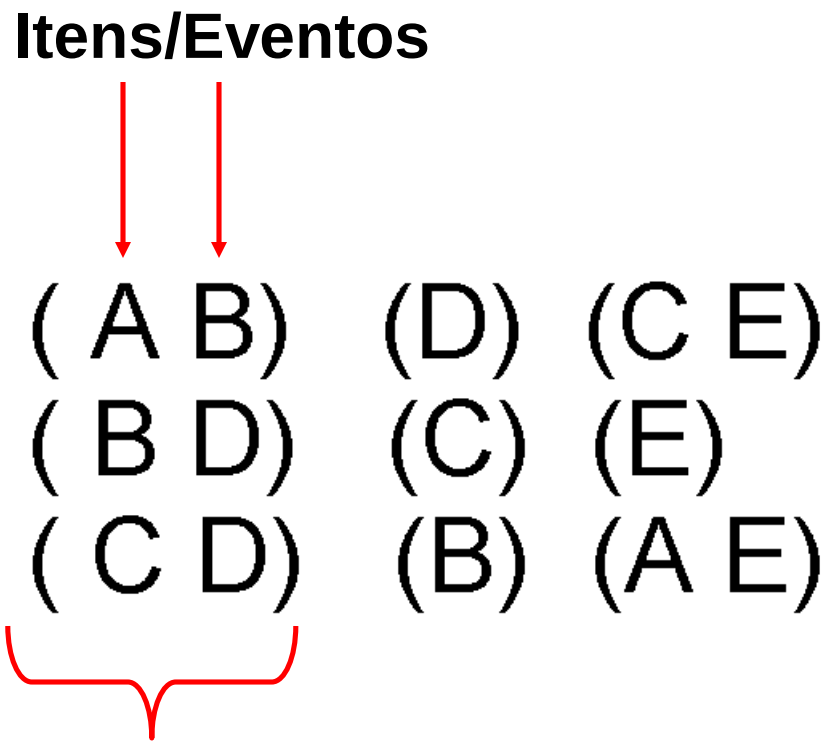
$$p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$$

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



Dados ordenados

- Sequências de transações



Dados ordenados

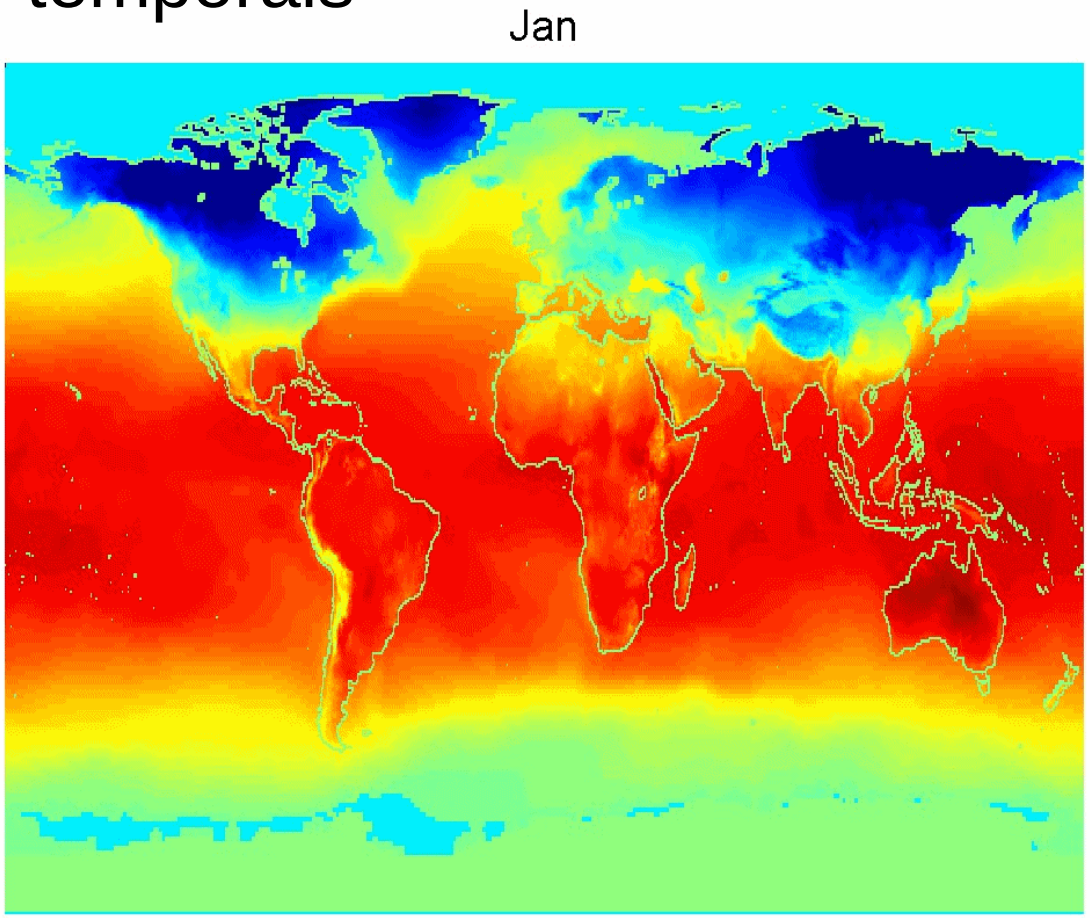
- Dados de sequência genômica

```
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

"O ponto mais importante é que a baixa qualidade dos dados é um desastre em desenvolvimento.

  - A baixa qualidade dos dados custa a empresa típica, pelo menos, dez por cento (10%) de receita; vinte por cento (20%) é provavelmente uma estimativa melhor."

Thomas C. Redman, DM Review, Augusto 2004

- 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

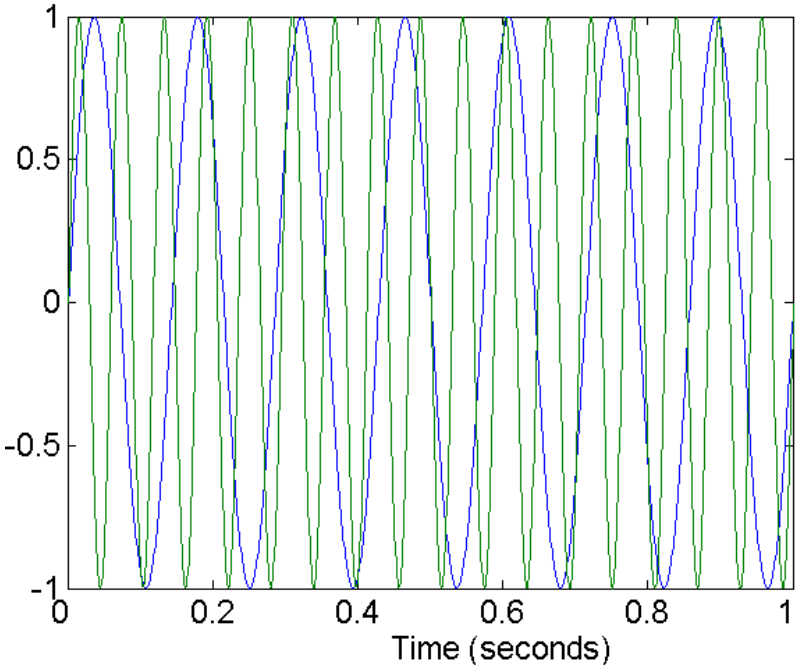
Qualidade dos dados ...

- Que tipos de problemas de qualidade de dados existem?
- Como podemos detectar problemas com os dados?
- O que podemos fazer sobre esses problemas?

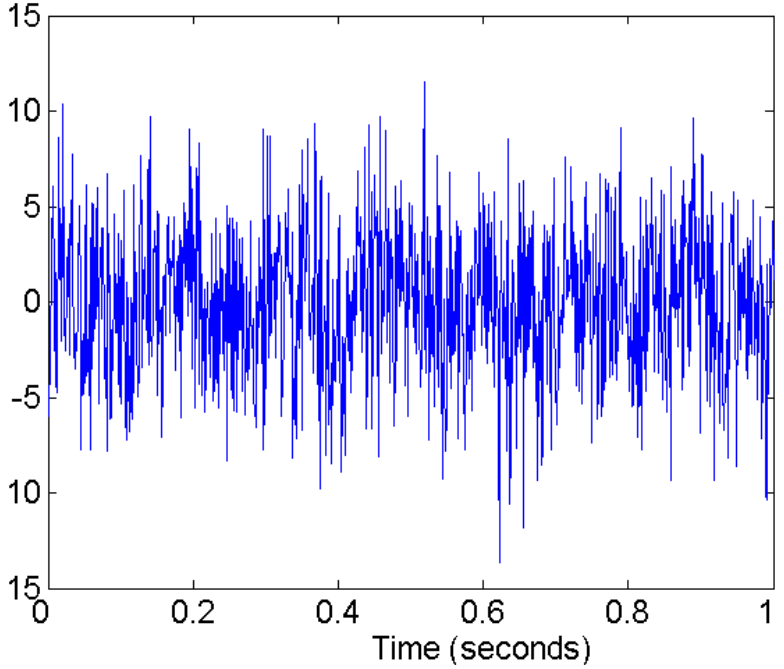
- Exemplos de problemas de qualidade de dados:
  - Ruído e outliers
  - Valores ausentes
  - Dados duplicados
  - Dados errados

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



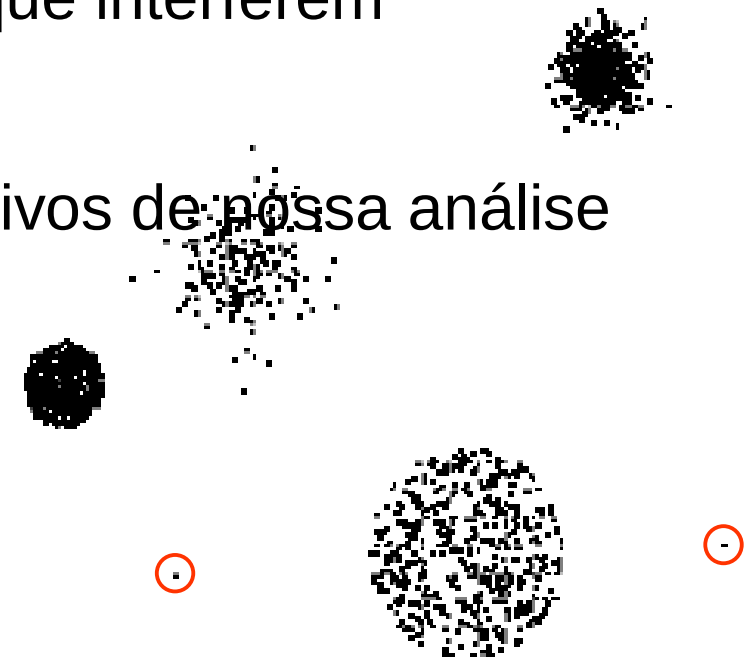
Duas ondas senoidais



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
- Causas?



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

Valores ausentes ...

- ... faltando completamente aleatoriamente (Missing Completely At Random - MCAR)
  - Falta de valor é independente de atributos
  - Preencha os valores com base no atributo
  - A analise pode ser imparcial em geral
- ... faltando aleatoriamente (Missing at Random - MAR)
  - Falta é relacionada a outras variaveis
  - Preencha os valores com base nos outros valores
  - Quase sempre resulta em viés na analise
- ... faltando não aleatoriamente (Missing Not at Random - MNAR)
  - Falta é relacionada a medições não observadas
  - Falta informativa ou não ignorável
- Não é possível saber a situação a partir dos dados

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?

Descrições estatísticas de dados

- Motivação
  - Para entender melhor os dados: tendência central, variação e dispersão
- Características de dispersão de dados:
  - Mediana, valor máximo e mínimo, quantis, outliers, variância, etc.
- Dimensões numéricas correspondem aos intervalos classificados
  - Dispersão de dados: analisados com múltiplas granularidades de precisão
  - Boxplot ou análise quantílica em intervalos ordenados
- Análise de dispersão em medidas computadas
  - Colocar medidas em dimensões numéricas
  - Análise de boxplot ou quantil no cubo transformado

Tendência central

- Média
  - n é tamanho da amostra, N é tamanho da população

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  - média ponderada

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - média truncada (trimmed)
    - retira-se percentual de valores extremos

Tendência central

- Mediana
  - valor do meio se quantidade impar, ou média dos valores do meio se quantidade par
  - atributos numéricos e ordinais
- Moda
  - valor mais frequente
  - uma moda: unimodal
  - bimodal, trimodal, multimodal
  - atributos numéricos, ordinais e nominais
  - Formula empírica:

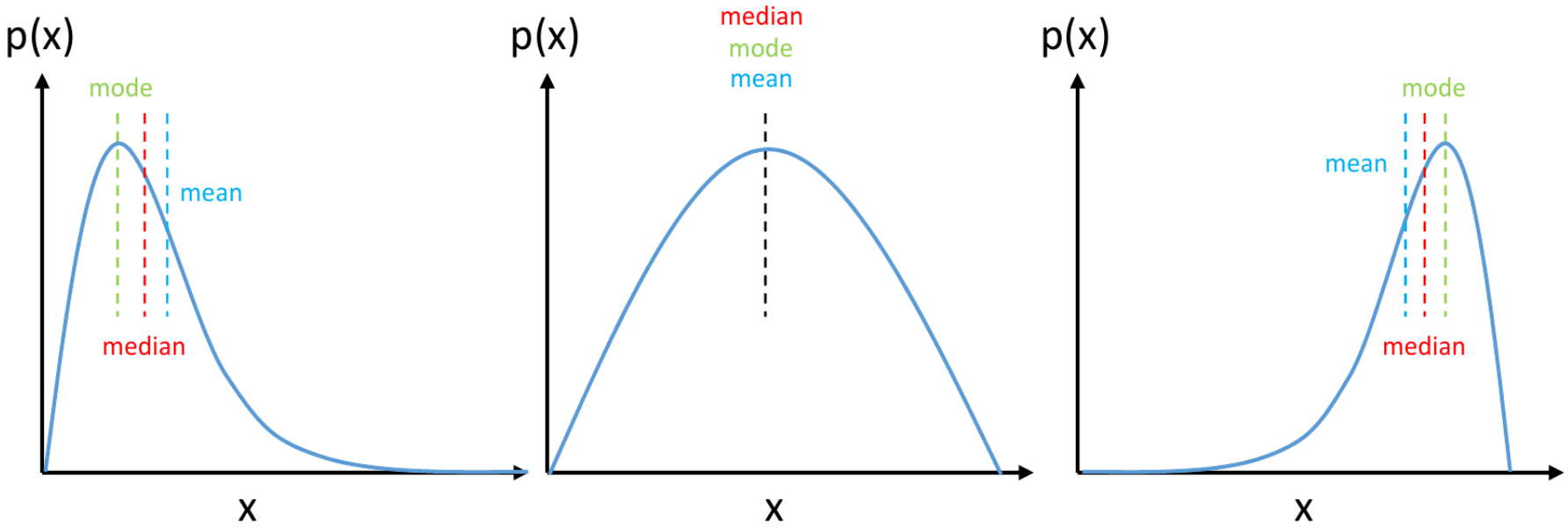
$$\text{mean} - \text{mode} \simeq 3 \times (\text{mean} - \text{median})$$

Tendência central

- Exemplo: Qual a média, mediana e moda destes dados?
  - 2, 3, 5, 5, 6, 6, 6, 7, 7, 9, 10
  - 1, 2, 2, 3, 3, 3, 5, 6, 11, 13, 17
  - 1.2, 1.4, 1.5, 1.8, 10.2

Dados simétricos e assimétricos

- Mediana, média e modo de dados simétricos, positivamente e negativamente assimétricos



Positivamente assimétrico
modo < mediana

Simétrico
modo = mediana

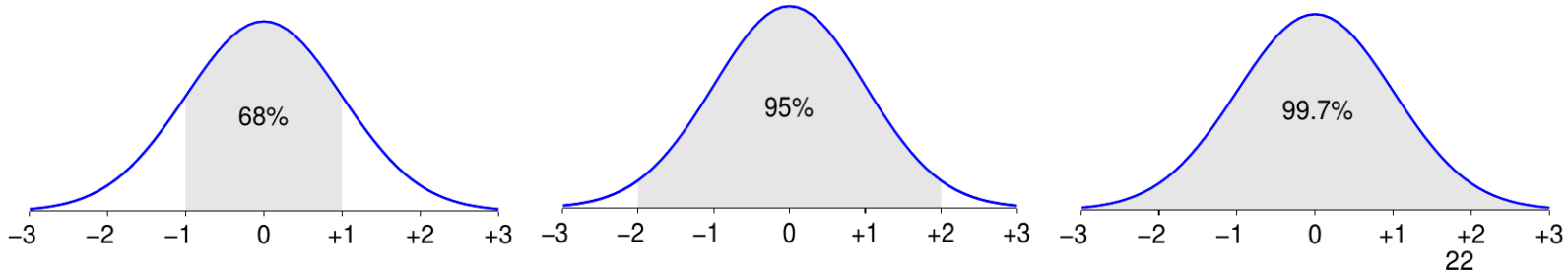
Negativamente assimétrico
modo > mediana

Medindo a dispersão de dados

- Variação e desvio padrão
  - Variação
  - O desvio padrão σ é a raiz quadrada da variância σ^2

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \mathbb{E}[x] \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

- A curva de distribuição normal
  - De μ – σ a μ + σ: contém cerca de 68% das medidas
  - De μ – 2σ a μ + 2σ: contém cerca de 95% de medidas
  - De μ – 3σ a μ + 3σ: contém cerca de 99,7% medidas



Dispersão de dados

- Quantil
  - Pontos que dividem dados ordenados em q subconjuntos de tamanho igual
  - Cada subconjunto é um q-quantil, teremos (q-1) q-quantis
  - q=100: os 100-quatil são percentis
  - q=4: os 4-quantil são quartis

Dispersão de dados

- Quartil, outliers e boxplots
  - Quartil: Q1 (percentil de 25%), Q3 (percentil de 75%)
  - Distância inter-quartil: IQR = Q3 – Q1
  - Resumo de 5 valores: min, Q1, mediana, Q3, max
  - Boxplot: final da caixa são os quartis, mediana é marcada, além de bigodes e outliers
  - Outlier: usualmente, um valor maior/menor que 1.5*IQR